[Vision Paper] PRObot: Enhancing Patient-Reported Outcome Measures for Diabetic Retinopathy using Chatbots and Generative AI

Maren Pielka*†§, Tobias Schneider*†§, Jan Terheyden‡, Rafet Sifa*†

*Fraunhofer IAIS, Sankt Augustin, Germany

†University of Bonn, Bonn, Germany

‡Universitätsklinikum Bonn, Bonn, Germany

§Equal contribution

Maren.Pielka@iais.fraunhofer.de

Abstract—We present an outline of the first large language model (LLM) based chatbot application in the context of patientreported outcome measures (PROMs) for diabetic retinopathy. By utilizing the capabilities of current LLMs, we enable patients to provide feedback about their quality of life and treatment progress via an interactive application. The proposed framework offers significant advantages over the current approach, which encompasses only qualitative collection of survey data or a static survey with limited answer options. Using the PROBot LLM-PROM application, patients will be asked tailored questions about their individual challenges, and can give more detailed feedback on the progress of their treatment. Based on this input, we will use machine learning to infer conventional PROM scores, which can be used by clinicians to evaluate the treatment status. The goal of the application is to improve adherence to the healthcare system and treatments, and thus ultimately reduce cases of subsequent vision impairment. The approach needs to be further validated using a survey and a clinical study.

Index Terms—Machine Learning, Large Language Models, Natural Language Understanding, Patient-Reported Outcome Measures, Ophthalmology

I. INTRODUCTION

Monitoring chronic eye diseases is increasingly based on data from imaging exams, functional tests and self-reports of patients. While treatment decisions in medicine depend on all three columns, quantitative data are often available only for imaging biomarkers and functional assessments, while symptoms and health behaviors are mostly assessed qualitatively. Structured questionnaires such as patient-reported outcome measures (PROMs), e.g. the National Eye Institute Visual Function Questionnaire 25 (NEI-VFQ-25, table I) [14], [15], are increasingly used to foster patient-centered care and provide quantitative metrics of patient-reports to clinicians. However, they have been developed in the context of clinical trials and are therefore highly static and not necessarily userfriendly. In PROM questionnaires, patients are asked questions e.g. about their physical and mental well-being, which are to be answered with a score on a numerical scale.

This work was accepted into the proceedings of the IEEE Big Data 2024 conference. The copyright lies with IEEE.

Question

How much of the time do you worry about your eyesight?

At the present time, would you say your eyesight using both eyes (with glasses or contact lenses, if you wear them) is excellent, good, fair, poor, or very poor or are you completely blind?

How much difficulty do you have driving in difficult conditions, such as in bad weather, during rush hour, on the freeway, or in city traffic?

How much pain or discomfort have you had in and around your eyes (for example, burning, itching, or aching)?

How much difficulty do you have reading ordinary print in newspapers?

Because of your eyesight, how much difficulty do you have visiting with people in their homes, at parties, or in restaurants?

TABLE I

EXEMPLARY QUESTIONS FROM THE NEI-VFQ-25 QUESTIONNAIRE FOR EYE DISEASE PATIENTS.

Due to the static and repetitive design of PROM questionnaires with little to no individualization, low adherence and loss to follow-up are common themes in real-world applications of PROMs in the context of chronic conditions. This impedes the high degree of patient empowerment that can be reached by implementing PROMs into health services. This issue becomes particularly relevant in the case of diabetic retinopathy, which is a prominent and increasing cause of global blindness. Advanced stages of diabetic retinopathy require invasive treatments such as regular injections into the eye, laser coagulation or surgery, which prevents blindness in the affected patients. However, a considerable proportion of people with diabetic retinopathy skips appointments or cancels the treatment altogether, which suggests that the communication between patients and care providers needs to be improved and patients need to be empowered at an earlier stage.

Given this status quo, and considering the impressive development of generative artificial intelligence (GenAI), and especially large language models (LLMs) in the last years, we came up with the vision of an interactive chatbot to replace the static PROM-questionnaires. Instead of lacking communication between patients and care providers or a

repetitive set of questions, patients can "speak" to the chatbot in real time and answer questions that are streamlined to their personal situation and background. In order to obtain numeric scores for measuring the progress of the treatment - similar to the conventional PROM framework -, another machine learning (ML) model is being trained to predict those based on the patient's responses. This approach is being validated by comparing the scores predicted by the ML model to those obtained from the original PROM questionnaire. We aim to cooperate with health providers and make our research available to patients via an application. According to the idea of merging the PROM framework with a chatbot interface, we call our approach "PRObot".

In the scope of this vision paper, we qualitatively evaluate the LLM approach, using synthetic data and chatbot interactions with GPT-40 [16]. The next steps involve setting up a survey, followed by a clinical study to get both simulated and actual patient data in order to train the ML interpreter model and quantitatively evaluate the results.

II. RELATED WORK

Traditionally, medical questionnaires such as PROMs are designed according to Likert-type scales [12]. This is a standard approach for surveys where the answers are to be given on a discrete scale. In this framework, single questions are referred to as "items", and the answer options usually correspond to the level of agreement or disagreement with a given statement.

There are a number of possible applications for artificial intelligence and machine learning in medicine, as well as ophthalmology specifically. The majority are image-based methods with a focus on detecting and classifying abnormalities [2], [18]. Using AI for interaction with patients - specifically PROMs - is a relatively under-developed area. There has been some previous work on using the outcome of PROMs to improve the therapy for certain diseases [11]. Improving or replacing the PROMs themselves by a generative AI approach has, to our best knowledge, not been attempted yet.

The fast development in the field of transformer-based language models [21], especially large generative pre-trained transformers (GPT) [3], [16], [17], has given rise to new possibilities for language modeling and conversational AI. While the most potent models such as GPT-40 are only accessible via commercial APIs, there are also a number of open-source solutions that achieve competitive results, such as Mixtral [7], LLaMa3 [20] or Phi-3 [8]. We aim to leverage the generative power of those models by deploying them in an interactive patient-chatbot setting, while providing background information about the traditional PROM items, as well as the patient's history in the prompt. In addition, we want to fine-tune smaller, encoder based models such as BERT [6] or XMLRoBERTa [5] for predicting the PROM scores based on the patient's answers.

III. METHODOLOGY

PROMs are established and well-researched instruments for collecting patient-recorded data. However, due to their static design, PROMs are limited in the information they can capture, as any information not covered by PROM items cannot be recorded. Designing and validating a PROM questionnaire is a complex and time-consuming iterative process. But even the items of an established and well-calibrated PROM can have a varying relevance due to the strong differences in humans. For example, in the NEI-VFQ-25 questionnaire [14], [15], the influence of people's eyesight on mobility in everyday life is primarily measured by questions relating to driving a car. However, these questions might be irrelevant for many individuals.

We propose a novel PROM framework that leverages the capabilities of LLMs to perform a dynamic and more personalized survey of patients. Instead of relying on static questions and predefined answer options, this framework is based on an interactive chat environment in which patients can respond with free text. Answers based on free text can convey much more information, enabling a more refined and personalized evaluation of responses. Additionally, this framework can perform personalized and dynamic questioning for individual patients, using general information about the patient, previous interviews, and even the current chat history to improve the relevance of questions. A chat environment is also capable of responding to the free text answers written by the patients. For example, a chatbot could react emphatically and optimistically to negative responses. If a patient reports improvements, a chatbot could react positively to help the patient value positive developments [19].

Using the NEI-VFQ-25 PROM as an example, our goal is to test out a large variety of machine learning models and approaches to derive an optimal framework for building LLM-based PROMS. We will report on the performance and benefits of certain approaches to derive recommendations for LLM-based PROMS and determine an optimal configuration of the PROBot framework.

A. Proposed Framework

The proposed framework consists of the three main components *Interpreter*, *Chatbot* and *Storage*, as visualized in Figure 1. The Interpreter is responsible for understanding and analyzing patient responses. Based on the results of the Interpreter and data from the Storage component, the Chatbot generates appropriate messages for the dialog with patients. The *Storage* component provides the available patient data such as past conversations or various patient information. The Interpreter uses encoder LLMs to extract meaningful vector embeddings of patient responses and the Chatbot uses LLM Prompts and Prompt Engineering to generate high-qualitative and creative text messages. Both components rely on the strong capabilities of modern LLMs.

The Interpreter consists of an *Encoder* LLM that is used to extract meaningful embeddings for free-text formulations, followed by a machine-learning based *Rater* which derives output

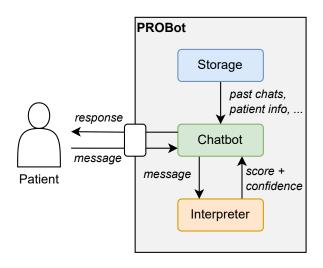


Fig. 1. Illustration of the PRObot framework

scores based on these embeddings. State-of-the-art Encoder LLMs, such as leading models for the MTEB Benchmark [4], will be fine-tuned using task-specific reference text data, which is collected in a survey. The Encoder LLMs can either be fine-tuned directly on the text data in a self- or unsupervised fashion. Or the entire Interpreter model can be trained end-to-end, which means training the Encoder LLM and Rater together, using the error of output scores as a training loss for the entire model.

Various ML models will be tested to derive output scores, using the NEI-VFQ-25 PROM as a reference. Different ML architectures such as MLPs or XGBoost will be compared. There are different reasonable model outputs, the most widely used being regression or classification: While regression considers the order of output classes, the output is continuous and typically there is no measure of confidence. Classification on the other hand produces discrete outputs. Using probabilities as output values can be interpreted as the model's confidence. But the ordering of classes is not considered in the training, which can harm the model's performance. But also more sophisticated approaches, such as Ordinal Classification [9] which models the order of classes, or Bayesian Regression [10] which models uncertainty for Regression, exist.

Another important design decision is the composition and sharing of components inside the framework architecture. The outputs of the model could either be a score for each PROM item, for each PROM subscale or the model could even directly calculate the final aggregated score. Additionally, the Embedder or Rater can be re-used for different PROM items/subscales or unique components can be trained for each one. Methods that map a reduced number of free text responses, such as one response per PROM subscale, directly to the subscale scores or the final PROM score have the potential to greatly reduce the required number of patient responses through the strong capabilities of LLMs. However, this also makes the model less transparent. Therefore, in

addition to performance, various configurations will also be evaluated with regard to explainability, as it is highly relevant for medical applications.

For the Chatbot component of the PROBot framework, powerful capabilities of modern LLMs are used to generate chat messages. An important decision when developing the Chatbot component is whether to implement it by training a task-specific LLM or by leveraging LLM Prompting with pre-trained models such as ChatGPT [16] or Llama-3 [20]. Training a new task-specific LLM would require a lot of clean data demonstrating the intended chatbot behaviour. This kind of data is difficult to obtain and the resulting model would have limited customizability. However, many pre-trained LLMs such as ChatGPT are intended for chatbot usage and therefore already fulfill many relevant requirements for our framework as LLM alignment is a major priority in the development of such models. Relevant aspects such as reliability, safety, fairness, resistance to misuse, explainability & reasoning, social norms and robustness are considered in LLM alignment [13]. By defining user prompts, LLMs become highly customizable as they allow for a precise description of the chatbot behaviour. Therefore, the chatbot's behaviour can also be easily optimized in the future. A prompt-based approach enables cost-effective reuse of the framework for other use cases as well. Therefore, we propose to use LLM Prompting for the Chatbot component.

B. Data Collection and Evaluation

To ensure high quality for the chatbot text generation and interpretation of patient responses, large amounts of clean data need to be collected to fine-tune and evaluate task-specific large language models. A large corpus of patient surveys will be acquired, based on the NEI-VFQ-25 PROM and free text formulations. This data allows the training and comparison of many kinds of machine learning models with respect to the broadly accepted NEI-VFQ-25 PROM. Each participant will fill out the NEI-VFQ-25 PROM, which can then be used for target scores when training and evaluating ML models.

The participants are asked to write free text formulations in different formats. To ensure that the time required to complete the survey is reasonable, each participant will only complete one of the formats. Participants may be asked to respond to each individual PROM item with a free text, or to respond to a question for each PROM subscale with a free text. These formats are useful as they allow a direct mapping from the free text formulations to the target scores of the NEI-VFQ-25 PROM. Other formats such as writing a single large free-text formulation or using other scales such as 10 quality-of-life domains are also possible. However, these may be more difficult to map to the target scores of the NEI-VFQ-25 PROM. In this case, the participants will provide a score for each written free text, which can also be used as a target for the training and evaluation of models.

The collected data will be used to train ML models to learn the mapping from free-text responses to scores such as the individual PROM items or the PROM subscales. By collecting free-text responses for individual PROM items as well as for PROM subscales, the ability of LLMs to derive meaningful scores for entire subscales with just one text can be analyzed.

To test the chatbot's text generation based on LLM prompting, interactive experiments will be conducted with individuals. By utilizing evaluation methods such as A/B testing, the performance of different LLM configurations can be compared. The iterative execution of this approach with an increasing number of participants allows to gradually evaluate and improve the chatbot's text generation capabilities. Conducting experiments with individuals is an expensive and time-consuming process. But the findings and data are very important for optimizing the chatbot for interaction with real people. While it is not reasonable to omit experiments with humans, it has been shown that LLMs can be utilized as strong user simulators [1]. Therefore, LLM-based user simulators could be used in earlier stages to optimize an initial LLM prompt for the Chatbot, which can be used as a baseline for the first experiments with real users.

The performance of the Interpreter will mainly be evaluated quantitatively by analyzing the error of predicted scores compared to their NEI-VFQ-25 target, using various regression or classification metrics, such as mean squared error, accuracy or F1-Score. Beside raw model performance, possible biases of the Interpreter such as age- or gender-biases should also be analyzed. On the other hand, evaluating the performance of the chatbot is not that straightforward. While quantitative evaluation, using explicit metrics such as user feedback or implicit metrics such as conversation length or abortion rate, can be used to gain insights about the quality of the text generation, qualitative analysis still plays an important role when evaluating the outputs for text generation tasks.

IV. QUALITATIVE RESULTS

To test the proposed chat generation approach, we have created a prototype for PRObot, which we call "PROTObot". In order to generate exemplary dialogues for an initial qualitative analysis, we simulate synthetic patients using LLMs. In turns, responses are generated for PROTObot and the simulated patient using two separate instances of GPT-4o. The PROTObot instance is assigned with asking the patient quality-of-life questions regarding their eyesight and address the patients responses to earlier questions with empathy. For this, the NEI-VFQ-25 questionnaire is provided as reference material. The instance for the simulated patients is tasked with responding to the messages of the PROTObot instance, while following an LLM generated persona. Table II and III show excerpts from generated dialogues with the synthetic patients Michaela Schmidt and Sandra Becker. The examples show that the simulation of patients results in the generation of exciting dialogues, which enables an initial evaluation of the proposed approach. One can see that PROTObot is able to generate individually suitable questions for the patients and react appropriately to the patient's responses. The generated questions are not just copied from the NEI-VFQ-25 PROM, but in most cases highly personalized questions that are tailored to each patient's available information. However, the scope of **PROTOBot**: Hello, I'm PROTOBot, here to support you with your eye condition. To better understand how your vision affects you, could you please tell me a bit about your profession, hobbies, and other activities that might be influenced by your eyesight?

User: I'm Michaela Schmidt. As a graphic designer, my blurry vision makes it really hard to work on detailed projects and color selections. I love photography and hiking, but focusing my camera and navigating rough trails have become increasingly difficult.

PROTOBot: Thank you, Michaela, for sharing that with me. It sounds like your profession and hobbies rely heavily on good vision. Let's delve a bit deeper to understand how your eye condition impacts different aspects of your life.

How often do your vision problems interfere with your ability to work on detailed projects and make accurate color selections?

User: My vision problems interfere with my work on detailed projects and color selection almost daily. It has become really hard to distinguish between subtle color differences and focus on fine details.

PROTOBot: That must be incredibly challenging, especially in a field where precision is so important. I'm truly sorry you're experiencing this difficulty.

How has your enjoyment of photography been affected by your blurry vision, particularly when trying to focus your camera or capture detailed shots?

User: It's been quite frustrating. I often struggle with focusing my camera, and capturing detailed shots has become very challenging.

PROTOBot: I can understand how disheartening that must be, given your passion for photography. It's tough when a loved hobby becomes a source of frustration.

TABLE II EXCERPT FROM EXEMPLARY DIALOGUE OF PROTOBOT AND A SIMULATED USER USING GPT-40

the questions remains similar to the reference questions. In the dialogue with Michaela, PROTObot demonstrates a great amount of empathy. Since she mainly reports on frustrating difficulties in her everyday life, this is very appropriate. PROTObot also shows a lot of empathy and understanding in the dialogue with Sandra. In the later messages, Sandra reports on techniques for dealing with her problems more positively. Here, PROTOBot highlights the positive aspects about these techniques, but without disregarding her underlying issues. Despite the strong empathy that PROTObot demonstrates, the interactions do not come across as overly solicitous, but rather natural. However, it should be noted that this can be perceived differently depending on the individual and should be evaluated further.

V. CONCLUSION AND OUTLOOK

We presented the outline for an AI-driven approach to improve the quality and usefulness of patient-reported outcome measures for diabetic retinopathy. Next steps involve conducting a survey, as well as a clinical study to collect data for training and validation. On this basis, we want to further refine our approach and make it accessible to health care providers and patients.

Further development includes validating the ideas developed in this vision paper, i.e. training and evaluating ML models for the chatbot setup and the score prediction. If the approach proves successful for diabethic retinopathy, it could easily be extended to other chronic diseases and prevention use cases, **PROTOBot**: Hello, I'm PROTOBot, here to support you with any questions you might have related to your eye health.

To better assist you, could you please tell me more about yourself? Specifically, your profession, hobbies, and other activities that may be affected by your eyesight.

User: Hello, I'm Sandra Becker. I'm a teacher, which means I spend a lot of time grading student work. I love painting and practicing yoga in my free time. My cataracts make my vision cloudy, so I have to focus on larger, simpler motifs in my artwork, and yoga helps me maintain my balance.

PROTOBot: Hi Sandra, thank you for sharing more about yourself. It sounds like your cataracts have quite an impact on your daily life and hobbies.

Could you describe how your cataracts affect your ability to read and grade student work?

User: Reading and grading student work has become much more challenging due to the cloudiness in my vision. I often have to use brighter lighting and magnifying tools to see the text clearly.

PROTOBot: I'm sorry to hear that reading and grading student work has become more difficult for you, Sandra. It's admirable how you are adapting to these challenges to continue your work.

How does your vision affect your ability to enjoy painting? Are there specific techniques or tools you've had to change or implement?

User: The cloudiness from the cataracts makes it hard to see fine details, so I focus on larger, simpler motifs when I paint. I've also started using brighter colors to make the contrasts more noticeable.

PROTOBot: That sounds like a positive approach and a creative way to continue enjoying your painting despite the challenges. It's great to hear that you are finding ways to adapt.

Has your vision impacted your ability to practice yoga, particularly with maintaining balance or following along with yoga classes?

User: Yes, my vision has made it a bit more challenging to maintain my balance during yoga. However, by focusing on my other senses and practicing regularly, I've managed to adapt and still enjoy my sessions.

PROTOBot: It's wonderful to hear that you've found ways to adapt and continue with your yoga practice despite the challenges. Maintaining such a positive attitude is truly inspiring.

TABLE III EXCERPT FROM EXEMPLARY DIALOGUE OF PROTOBOT AND A SIMULATED USER USING GPT-40

such as oncology, nutrition, or smoking dehabituation. We find that LLMs and chatbots offer great opportunities to help patients as well as health care professionals in understanding and improving factors for treatment success.

DISCLOSURES

From our point of view, the following conflicts of interest with health care providers are possible: Bayer (Funding), Novartis (Funding, Speaker Fees), Roche (Funding), Okko (Consultancy).

ACKNOWLEDGMENT

The project was funded by the Federal Ministry of Education and Research (BMBF) under grant no. 01IS22094A WEST-AI.

This research has been partially funded by the Federal Ministry of Education and Research of Germany and the state of North-Rhine Westphalia as part of the Lamarr-Institute for Machine Learning and Artificial Intelligence.

REFERENCES

- Atheer Algherairy and Moataz Ahmed. Prompting large language models for user simulation in task-oriented dialogue systems. Computer Speech & Language, 89:101697, 2025.
- [2] David Biesner, Helen Schneider, Benjamin Wulff, Ulrike Attenberger, and Rafet Sifa. Improving chest x-ray classification by rnn-based patient monitoring. In 2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA), pages 946–950, 2022.
- [3] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Proc. NeurIPS*, volume 33, pages 1877–1901, 2020.
- [4] Hongliu Cao. Recent advances in text embedding: A comprehensive review of top-performing methods on the mteb benchmark, 2024.
- [5] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019.
- [6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. CoRR, abs/1810.04805, 2018.
- [7] Albert Q. Jiang et al. Mixtral of experts, 2024.
- [8] Marah Abdin et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024.
- [9] Eibe Frank and Mark A. Hall. A simple approach to ordinal classification. In *European Conference on Machine Learning*, volume 2167, pages 145–156, 08 2001.
- [10] Michael M. Goldstein. Bayesian analysis of regression problems. Biometrika, 63:51–58, 1976.
- [11] Hadas Lewy, Refael Barak Barken, Tomas Karpati Karpati, and Amir Glik.
- [12] Rensis Likert. A technique for the measurement of attitudes. Archives of Psychology, 140:1–55, 1932.
- [13] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. Trustworthy llms: a survey and guideline for evaluating large language models' alignment, 2024.
- [14] Carol M. Mangione, Paul P. Lee, Peter R. Gutierrez, Karen Spritzer, Sandra Berry, Ron D. Hays, and for the National Eye Institute Visual Function Questionnaire Field Test Investigators. Development of the 25list-item National Eye Institute Visual Function Questionnaire. Archives of Ophthalmology, 119(7):1050–1058, 07 2001.
- [15] Carol M. Mangione, Paul P. Lee, Jennifer Pitts, Peter Gutierrez, Sandra Berry, Ron D. Hays, and for the NEI-VFQ Field Test Investigators. Psychometric Properties of the National Eye Institute Visual Function Questionnaire (NEI-VFQ). Archives of Ophthalmology, 116(11):1496– 1504, 11 1998.
- [16] OpenAI. Gpt-4 technical report. arXiv:2303.08774, 2023.
- [17] A Radford. Improving language understanding by generative pretraining. arXiv, 2018.
- [18] Helen Schneider, Priya Priya, Aditya Parikh, Christian Bauckhage, and Rafet Sifa. The potential of acknowledging the unknown: Single positive multi-label learning in medical image processing. In *Medical Imaging* with Deep Learning, 2024.
- [19] MD Vera Sorin, MD Danna Brin, MD Yiftach Barash, MD Eli Konen, MD PhD Alexander Charney, MD Girish Nadkarni, and MD Eyal Klang. Large language models (Ilms) and empathy - a systematic review. In medRxiv, 2023.
- [20] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.
- [21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.