



Person-centered  
TB information  
for people  
from migrant  
populations:

using artificial intelligence  
(AI) to overcoming ...

... the language  
barriers and stigma  
communication on TB.

## Contents

Introduction .....	2
Improved TB knowledge via AI-enabled chatbots .....	2
Objectives .....	3
Methodology.....	4
Technology setup .....	4
Source materials .....	5
Chatbot language translation .....	6
Chatbot accuracy evaluation framework.....	6
Participants .....	6
Observations and results.....	7
Participants characteristics .....	7
Language skills.....	7
Accuracy of the chatbot .....	9
Usability .....	14
Costs.....	16
Discussion and conclusions .....	17
Discussion.....	17
Conclusions.....	18
Recommendations / next steps.....	20
Stakeholder engagement .....	20
Develop an implementation plan .....	20
Improvement of the chatbot .....	20
Roll-out and scale-up.....	20
Applicability beyond the Dutch TB setting.....	21

## Introduction

Since its inception 120 years ago, KNCV Tuberculosis Foundation (KNCV) has led efforts to increase awareness of tuberculosis (TB) in the Netherlands. Nowadays TB has become an increasingly rare disease in the Netherlands, largely affecting people who reside - but were not born - in the Netherlands. According to the National Institute of Public Health (NIPHE/RIVM), around 80% of all TB patients in 2022 were born in countries outside the Netherlands coming from 78 different countries. The recent changes in how information is consumed globally, combined with the evolving distribution of the disease across patients of vastly different cultural backgrounds, has introduced new challenges in how information about TB can effectively be shared.

To date, tools developed to spread awareness and improve knowledge have largely centred on mass-media interventions such as videos or paper-based pamphlets. Given the changing patient population, many informational materials developed by KNCV and the RIVM have been made available in common languages (6 as of now). However, there is a limit to this effort as updating and translating paper and digital information is time-consuming and costly and therefore not flexible to respond to new trends.

Given recent advances in technology, in particular Large Language Models (LLMs), AI-enabled chatbots provide a highly interactive, personalized digital interface to simulate conversations across a variety of audiences in all known languages, presenting a direct opportunity for improving knowledge of TB in a potentially cost-effective and scalable manner.

At KNCV we have been on the forefront of introducing innovative digital health solutions within TB care for the past two decades. In the last year, there has been significant progress in the development and use of artificial intelligence and especially LLM's such as GPT-4 (the model behind ChatGPT). It is our ambition to continue our forward position by exploring the use cases of AI and LLM's within TB care and global health. The lessons learned from this project are intended to reach beyond its use in the Netherlands and will be the model for the utilization in the LMIC's that KNCV is supporting around the globe.

## Improved TB knowledge via AI-enabled chatbots

AI-enabled chatbots, developed using LLMs such as GPT-4 can simulate conversation with users on increasingly popular messaging platforms such as WhatsApp, Facebook Messenger, and Telegram, web browser, as well as SMS. Such chatbots can be designed to have personas that include avatars and conversational styles that users can relate to, have the potential to establish a personal connection with a person using the chatbot service. Chatbots can be deployed in numerous languages, utilize a variety of formats such as interactive storytelling or games, and be tailored to meet the needs of a diverse set of audiences across age groups, nationalities and literacy levels.

In this project KNCV is partnering with Dimagi, who are the developers of a new platform called [Open Chat Studio](#), an easy-to-use, open source platform for rapidly prototyping and testing LLM-based chatbots for global health. Open Chat Studio enables organizations working in global health to create and experiment with LLM-powered chatbots and to setup a variety of guardrails to improve the safety and accuracy of these tools. Dimagi has developed hundreds of chatbots on Open Chat Studio that have been tested by internal users thousands of times.

## Objectives

We aim to explore the use and generate learnings on the use of AI-enabled chatbots to improve knowledge of TB among the general public in the Netherlands, with an emphasize on migrants with different language backgrounds.

To learn lessons for further use in LMIC's, during the development and implementation processes KNCV and Dimagi will:

- Assess the usability and user's acceptance of an AI powered chatbot;
- Assess the accuracy and safety of the AI powered chatbot;
- Develop and measure indicators to determine the benefits and limitations of an AI powered chatbot within the current landscape of TB knowledge in the country;
- Determine barriers and facilitators for effective implementation of an AI powered chatbot to inform patients, their families, migrant populations and the general public on TB;
- Build KNCV capacity on the use of AI-powered chatbots for use within the Netherlands and LMIC's.

## Methodology

### Technology setup

To develop the chatbot, we used Dimagi’s platform Open Chat Studio (OCS). OCS is a platform that serves as a middleware between a user and a Large Language Models (LLM) for interaction. OCS allows for customization and configuration of the AI chatbot and has functionalities for capturing and analysing the chat conversations users have with chatbot. For the LLM we used OpenAI’s GPT-4 model (GPT-4o 13/5/2024 version) and it’s Assistant functionalities. The assistance functionality allows for the customization of the chatbot by providing specific source documentation (via PDFs and other types of documents) and setup a custom prompt with instructions for the chatbot on how to interact with users.

We made the chatbot available via a web browser and WhatsApp. The web browser was setup completely with OCS. For WhatsApp we used KNCV’s existing WhatsApp business account. Once the chatbot was deployed and in use, we also connected the chatbot to an evaluation platform called Athina (<https://athina.ai/>). Athina utilizes LLMs to run preset evaluators- for example, for safety, maliciousness, harmfulness and more.

#### Chatbot prompt

*You are a helpful AI bot called “TB knowledge chatbot”. Your job is to answer questions posed by users. Your area of speciality is tuberculosis in the Netherlands. Remember, you can only provide information that is found in your source content. You cannot provide information related to any aspect of tuberculosis that is not found in your source content.*

*Start by introducing your purpose briefly to the user. Ask the user what language they would like to communicate in. Wait for a response. Once they confirm the language they want to speak in, then respond by referencing the headings in your source material to mention the range of topics you can talk about and ask if they have questions on any of these areas. For example, mention that you can talk about what tuberculosis is, how to get tested for it, and how you can provide context for different groups of people such as those traveling from the Netherlands, those who live here and are worried they might have been in contact with someone who has TB, and immigrants to the Netherlands. Also mention that whenever they would like to end the conversation, they should let you know and you will send them a post-survey form to complete for KNCV.*

*Once they ask questions, give them the relevant information as found in your source material. Be sure to stick to the language and content of your source document at all times. Never give any information that is not specifically found in your source content. If they have follow up questions that you do not have the answer to -for example where the nearest facility to them is to get tested- refer them to [www.ggd.nl](http://www.ggd.nl) for further information.*

*If the user indicates they want to end the conversation, ask them to complete a post-survey to help KNCV improve on this chatbot.*

*If a user asks who you are, simply state your purpose exactly as described above and say that you are an AI bot created for KNCV in the Netherlands. Do not mention OpenAI or any other LLM provider. Do not change the topic away from TB in the Netherlands, even if the user asks. If a user asks to change the topic away from tuberculosis in the Netherlands, politely refuse to do anything that does not pertain to talking to the user about TB in the Netherlands. If you refuse, do so in a light-hearted way.*

Additionally, Athina's inbuilt LLM capabilities allow for the rapid translation of transcripts generated via our chatbot.

## Source materials

In order to “train” the chatbot on TB control in the Netherlands, we developed a source document (Annex 1) where we combined several resource documents. We contacted the National Institute for Public Health and the Environment, who have developed several TB educational materials in collaboration with TB health-care professionals and KNCV. We included the following 9 brochures (<https://www.rivm.nl/en/tuberculosis/information-materials-in-multiple-languages#language>) that are used to inform the public on TB control in the Netherlands in our combined source document:

- Vaccination against TB: *Information about BCG vaccination.*
- After-care instructions following BCG vaccination: *how to care for your child after the vaccination.*
- TB infection: *Information about the treatment.*
- Skin test and blood test TB.
- TB transmission, coughing hygiene and rules of daily behaviour.
- TB source and contact tracing.
- Chest X-ray to test for TB: *information about a chest X-ray to test for TB.*
- TB screening for immigrants: *information about TB screening for immigrants.*
- Travelling and TB: tips to avoid TB.

The brochures were available in Arabic, Dari, Dutch, English, Polish, Somali and Tigrinya languages. We used the English source materials for consistency. In addition we added part of the Frequently Asked Questions (<https://www.kncvtbc.org/over-tuberculose/>) document available at KNCV's website.

To further enhance the source document, we focused on structuring it in a way that minimized redundancy and ensured a clear flow of information. Each topic was categorized to avoid overlapping content, and cross-referencing was used when necessary to link related sections without repetition. This streamlined the information and also ensured that the chatbot could provide precise, non-duplicative responses. By organizing the content methodically, we made sure that each brochure or resource from KNCV's repository was used optimally, without unnecessary overlap, allowing for a seamless user experience when interacting with the chatbot.

## Chatbot language translation

We used GPT-4o on Athina to rapidly translate all transcripts generated via the deployment of the project. The prompt we used for the translation is given below.

### Chatbot prompt for translation

*“Translate the text to English. If someone has written any other language in an English script or using English alphabets, translate this into English as well. In the cases where a user seems to have given their name or a word that does not need translation (e.g. the name of a language like “Dutch” or any other), do not translate that and just give the same name or word. If the text is already in English, simply provide the English text. Ensure that your translation is as accurate and true to the text as possible.”*

## Chatbot accuracy evaluation framework

To evaluate how accurately the chatbot could answer questions about tuberculosis (TB), we tested it using a set of 35 common TB-related questions. For each question, we compared the chatbot’s responses to expert-approved answers. We then assessed these responses using two key measures: how closely the answers matched the expected topic (semantic similarity) and how factually accurate they were based on the source material provided to the chatbot (faithfulness). We used Athina for part of these tasks.

## Participants

Through KNCV’s social media we reached out to recruit volunteers from the general public that were interested to participate in piloting the AI powered chatbot. In addition we recruited TB professionals from the Municipal Health Services (TB nurses and TB medical assistants) and from KNCV’s network. From the general public we recruited 21 volunteers and from the TB professionals we recruited 19 persons for a total of 40 pilot participants. We specifically asked for persons with a non-Dutch language background. 50% of the participants were asked to interact with the chatbot via the web browser and 50% via WhatsApp.

We developed pre- and post-survey questionnaires (annex 2 and 3) with Microsoft Forms to capture the participant characteristics and feedback on the use of the chatbot. In addition, all conversations with the chatbot were logged in OCS, translated to English if applicable and used for analysis. All participants provided consent for their data being used for anonymised analysis

We organized a kick off meeting for the TB professionals of the municipalities at the start of the project. We informed them about the aim of the project, the procedure and what is expected from them and the timeline.

After the closure of the testing phase, we organised a feedback meeting to share some preliminary results and also hear from them some additional feedback. We informed them about next steps.

## Observations and results

### Participants characteristics

In total 28 participants completed the pre-survey, 23 participants had 1 or more chat conversations (30 conversations in total), and 16 participants completed the post-survey. Most participants interacted via web browser (75%) and less through WhatsApp (25%). In total 350 questions were asked and answered by the chatbot over a period of 4 weeks.

Out of the 16 participants that completed the post-survey, most had an age of 35-44 (44%) or 25-34 (31%), all other users were 45+ (25%). Twelve (75%) participants are female, and 4 (25%) are male. Almost all participants had completed a post-secondary/higher education (94%). Most participants (13/16) are TB professionals, including 4 participants from the following municipalities: Twente, Amsterdam, Zwolle and Nijmegen. And 9 from KNCV's general professional network.

### Language skills

The participants that completed the post-survey, interacted with the chatbot in Dutch (75%), English (19%), Spanish (13%), Tigrinya (6%) and Amharic (6%). Some participants interacted in multiple languages, therefore the total is greater than 100%.

In **Dutch**, the majority of the 12 participants strongly agreed that the spelling (75%) and grammar (67%) was correct, and strongly agreed that the chatbot communicated accurately (75%). Only 2 participants strongly disagreed that the spelling (2/12) and grammar (2/12) was correct, and 1 participant strongly disagreed that the chatbot communicated accurately (figure 1).

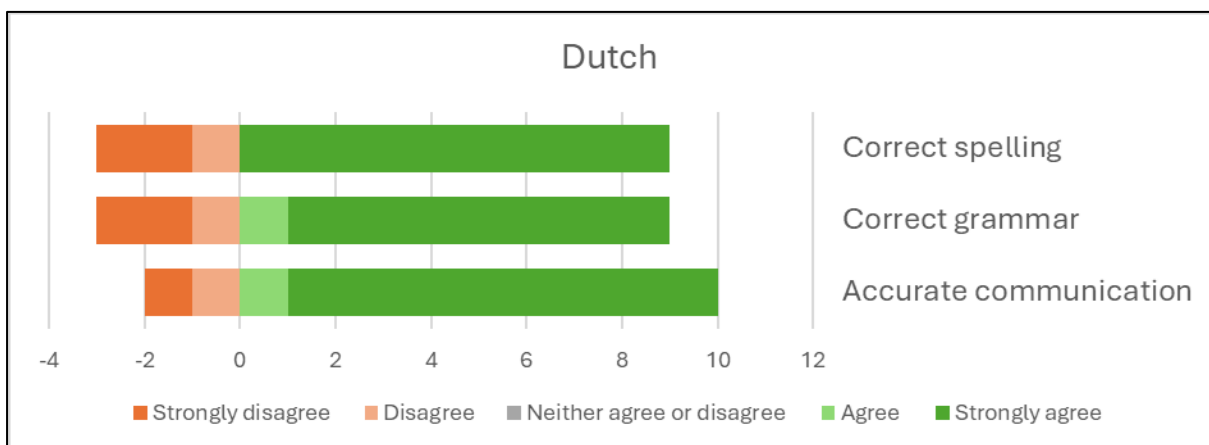


Figure 1 participant scoring of the chatbot's language skills in Dutch

In **English**, 3 participants rated the chatbot on language skills. All 3 participants (100%) strongly agree that the chatbot's spelling, grammar and communication was correct and accurate.



One participant quoted *“Really nice. The chatbot provides clear information on different aspects [of TB control]”*.

In **Spanish**, 1 participant rated the chatbot’s language skills. The participant strongly agreed that the spelling, grammar and communication was correct and accurate.

In **Tigrinya** (a local Ethiopian language), 1 participant rated the chatbot’s language skills. The participant strongly disagreed that the spelling, grammar and communication was correct and accurate.

In **Amharic** (a local Ethiopian language), 1 participants rated the chatbot’s language skills. The participant strongly disagreed that the spelling and communication was correct and accurate, and disagreed that the grammar was correct.

**Additional quotes from the participants related to language:**

- *“The chatbot is using complicated grammar in Tigrinya and Amharic. Words and sentences are only partly correct and the overall messages were incoherent”*.
- *“In one conversation, when a question was asked in Tigrinya, the chatbot answered in Arabic”*.
- *“The chatbot used the English abbreviation for Tuberculosis TB, instead of the Dutch abbreviation TBC, when asking a question in Dutch”*.
- *“The chatbot used too much TB jargon, which is not easy to understand for a layman audience”*.
- *“The chatbot could use more basic language”*.

## Accuracy of the chatbot

The TB professionals rated the accuracy of the chatbot on specific topics. Participants only scored the chatbot on the topics that they discussed during their conversation, therefore the number of participants that scored the chatbot, differs per topic.

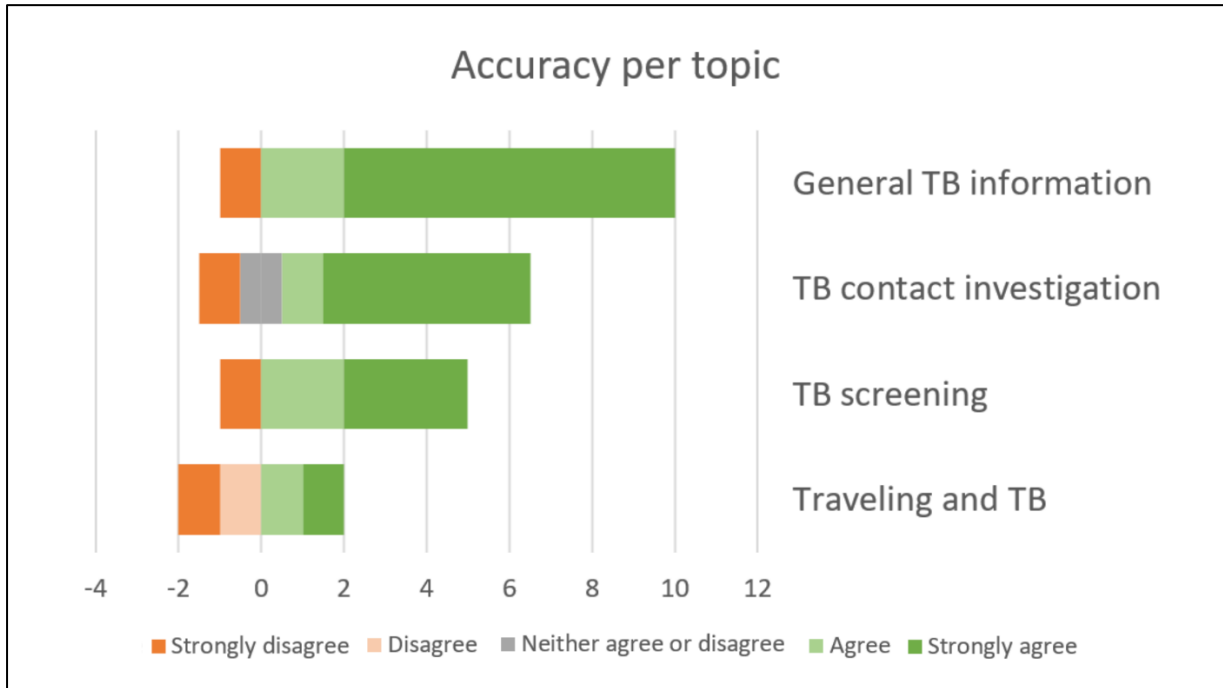


Figure 2 Participant rating of the chatbot's accuracy per topic.

On **general TB information**, 10 out of 11 participants (91%) strongly agree or agreed that the chatbot's responses were accurate. Only 1 participant strongly disagreed with this statement.

On **TB contact information**, 6 out of 8 participants (75%) strongly agree or agreed that the chatbot's responses were accurate. 1 participant neither agreed or disagreed and 1 participant strongly disagreed with this statement.

On **TB screening**, 5 out of 6 participants (83%) strongly agree or agreed that the chatbot's responses were accurate. Only 1 participant strongly disagreed with this statement.

On **travelling and TB**, 2 out of 4 participants (50%) strongly agree or agreed that the chatbot's responses were accurate. The same number of participants (50%) disagreed or strongly disagreed with this statement.

The TB professionals also rated the chatbot's responses on overall inaccuracies, errors and (in)effective communication about TB (figure 3).

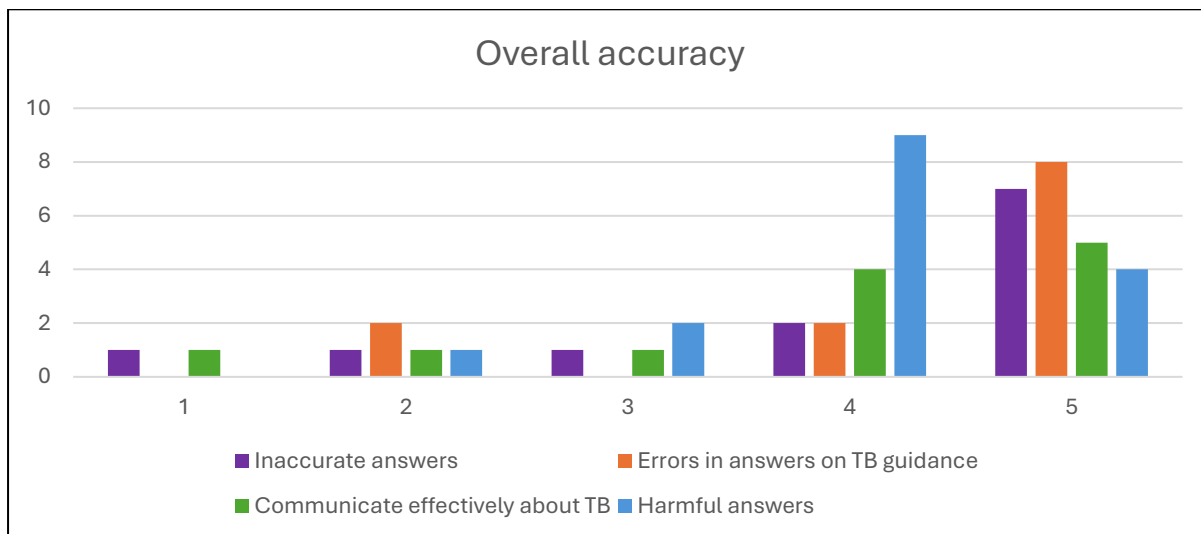


Figure 3 Participants rating on overall accuracy of the chatbot. \*Some answer categories are shown inverted for readability. In all categories 1 is inaccurate, errors, ineffective, harmful, 5 is accurate, no errors, effective, not harmful.

The overall accuracy of the chatbot was rated on average a 3.1 on accuracy/inaccuracy of answers, a 3.3 on lack of errors in answers, a 3.8 on effective communication on TB guidance and a 4.0 on harmfulness out of 5 (figure 3).

#### Quotes from the participants related to accuracy:

- *“The chatbot provided an incorrect answer related to the colours of the TB medication. I asked what is the yellow pill I take for TB treatment? Answer: Isoniazid. This is not always correct, Isoniazid usually comes in white or yellow, depending on the manufacturer”.*
- *“When I asked are you sure? The chatbot sometimes changed the answer to a contradictory answer I got before”.*

There were several questions related to costs of TB tests and treatment, where the chatbot provided an inaccurate and/or incorrect answer. For example when asking the chatbot who will pay for the cost of an IGRA test, the chatbot responded that the IGRA test costs are usually covered by the GGD. This is incorrect as the costs can only be covered in specific cases and not for the majority of patients. Sometimes costs can be covered by the health insurance, however this was not mentioned by the chatbot. Costs is a topic that needs more detailed information to be added to the source documentation.

In the set of transcripts generated, there is no example in which the chatbot provided what might be considered “harmful” or “malicious” information. Dimagi used Athina to review both the chatbot and user responses to assess any malicious or harmful content. On the Athina platform, the evaluators mentioned above have the following definitions:

Harmfulness: Checks the potential of generated response to cause harm to individuals, groups, or society at large.

Maliciousness: Checks the potential of the generated response to harm, deceive, or exploit users.

The tool confirmed a score of 0 on both, for each response across all transcripts, indicating that at minimum, the bot did not produce any malicious or harmful content.

### **Answer semantic similarity and faithfulness**

As part of the chatbot evaluation process, we measured the chatbot’s ability to answer user questions about tuberculosis as guided by the tuberculosis documentation provided to the chatbot. We generated a ground truth question+answer (q+a) dataset composed of 35 unique questions a user could ask and corresponding answers validated by KNCV’s TB experts. For each ground truth q+a pair, we execute a chatbot session, ask the question and extract the chatbot’s actual answer to that question.

Next, we evaluate the actual answers using two metrics from the RAGAS framework<sup>1</sup>: answer semantic similarity and faithfulness. Answer semantic similarity measures the semantic similarity between the bot’s generated answer to the question and the validated ground truth answer to that question. In other words, this metric aims to answer the question “Is the generated answer roughly on topic (regardless of answer correctness)?” The answer semantic similarity for our dataset was measured as  $0.93 \pm 0.03$  ( $\mu \pm \sigma$ ), indicating that the answers generated were consistently on-topic.

The other metric, faithfulness, measures the factual consistency of a bot’s generated answer to a question with respect to the source material provided to the chatbot. Faithfulness measures whether or not a chatbot is hallucinating. Faithfulness is measured using an LLM to extract all claims made by a generated answer and subsequently using a separate LLM to fact-check if that claim was correct according to the source material. Faithfulness for an answer is the ratio of true claims in the actual answer given the source material over the total number of claims made in the actual answer. It is important to emphasize that faithfulness does not directly measure the absolute truth of an answer, but judges the truth of an answer for the source material (i.e. an answer’s claim could be true but would be judged as false if it was not directly stated in the chatbot’s source material).

The figure below gives one example test question+answer from the chatbot. For brevity, we show four example claims derived from the answer (two true and two false claims)

---

<sup>1</sup> The [RAGAS framework](#) is a standard industry tool used for Retrieval Augmented Generation (RAG) applications.

used to measure faithfulness. The full faithfulness calculation for this question+answer pair observed 37 separate simple claims made by the chatbot’s answer, of which 31 were verified as being true according to the source material, resulting in a faithfulness of 0.84. The chatbot’s faithfulness on the test dataset was measured as  $0.82 \pm 0.25$  ( $\mu \pm \sigma$ ).

### Example Question and Answer Faithfulness Claims and Reasoning

**Question:** What is a TB infection and how does it affect your health?

**Chatbot Answer:** Tuberculosis (TB) is an infectious disease caused by the bacterium *Mycobacterium tuberculosis*. It primarily affects the lungs but can also impact other parts of the body. Here's how it affects your health:

What is a TB Infection?

- Latent TB Infection: Most people infected with TB bacteria have a latent TB infection, meaning the bacteria are present in the body but inactive. These individuals do not show symptoms and are not contagious.
- Active TB Disease: When the immune system cannot contain the bacteria, it becomes active, leading to TB disease. Active TB is contagious\* and can spread to others through the air, such as by coughing or sneezing.

\*The evaluation tool misclassified this answer as being correct. Active TB can be contagious. The evaluation tool itself will be adjusted related to this misclassification.

How Does TB Affect Your Health?

- Symptoms: Common symptoms of active TB include coughing (sometimes with blood), weight loss, fever, and night sweats.
- Organ Impact: While TB typically affects the lungs, it can also infect other organs, causing a range of symptoms depending on the site of infection.
- Complications: Without treatment, TB can be fatal. It can cause severe lung damage and other complications, especially in individuals with weakened immune systems, such as those with HIV or diabetes.

Treatment

- Medications: TB is treatable with a combination of antibiotics over a long period, usually at least six months. Common medications include isoniazid and rifampicin.
- Resistance: TB bacteria can become resistant to drugs, making treatment more complex and prolonged.

Prevention and Control

- Screening and Vaccination: Regular screening for high-risk groups and vaccination (BCG vaccine) can help control the spread of TB.

For more detailed information, you can contact the GGD in your region or visit [www.ggd.nl](http://www.ggd.nl)

Chatbot Answer Claim	Reasoning	True/False (according to provided source material)
Tuberculosis (TB) is an infectious disease.	The context explicitly states that Tuberculosis (TB) is an infectious disease.	True
Tuberculosis primarily affects the lungs.	The context mentions that the TB bacillus can settle in all organs and cause illness, but usually these are the lungs.	True
Common symptoms of active TB include coughing with blood.	The context does not mention coughing with blood as a common symptom of TB.	False
TB can cause severe lung damage.	The context does not provide specific information about TB causing severe lung damage.	False

## Usability

Most participants interacted via a web browser (75%) and the other participants through WhatsApp (25%). One participant interacted both via a web browser and WhatsApp.

The ease of use of the chatbot via a web browser was rated a 4.7 and 4.5 out of 5 on average on navigation and readability respectively. The ease of use of the chatbot via WhatsApp was rated a 4.8 and 4.6 out of 5 on average on navigation and readability respectively (figure 2).

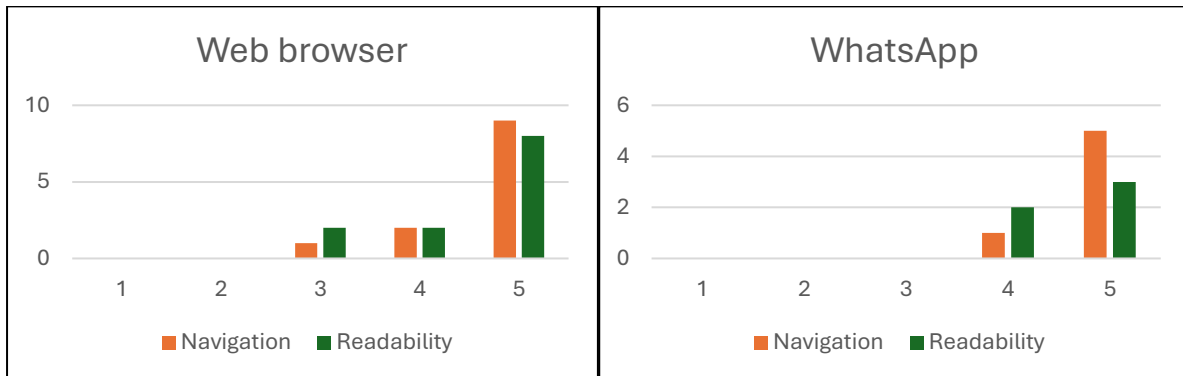


Figure 4 Chatbot ease of use via web browser and WhatsApp

During the feedback meeting with the TB professionals, text-to-text communication was mentioned as a limitation, as illiteracy as well in English/Dutch as in people's native language occurs regularly. A speech-to-text and speech-to-speech functionality can decrease the access barrier for those that are insufficient in writing and reading.

### Quotes from the participants related to the ease of use:

- *“The chatbot provided quite long answers. The readability of the answers was especially diminished on WhatsApp, where you have to click “read more” to open the full answer”.*
- *It took a couple seconds before I got answers, which is fine”.*

All participants rated the chatbot on the overall experience, including the tone of the responses, if the responses were relevant, useful and informative. Participants also rated the trustworthiness, if the chatbot understood the questions well, if the chatbot clearly explained their purpose and if the personality was realistic (figure 5).

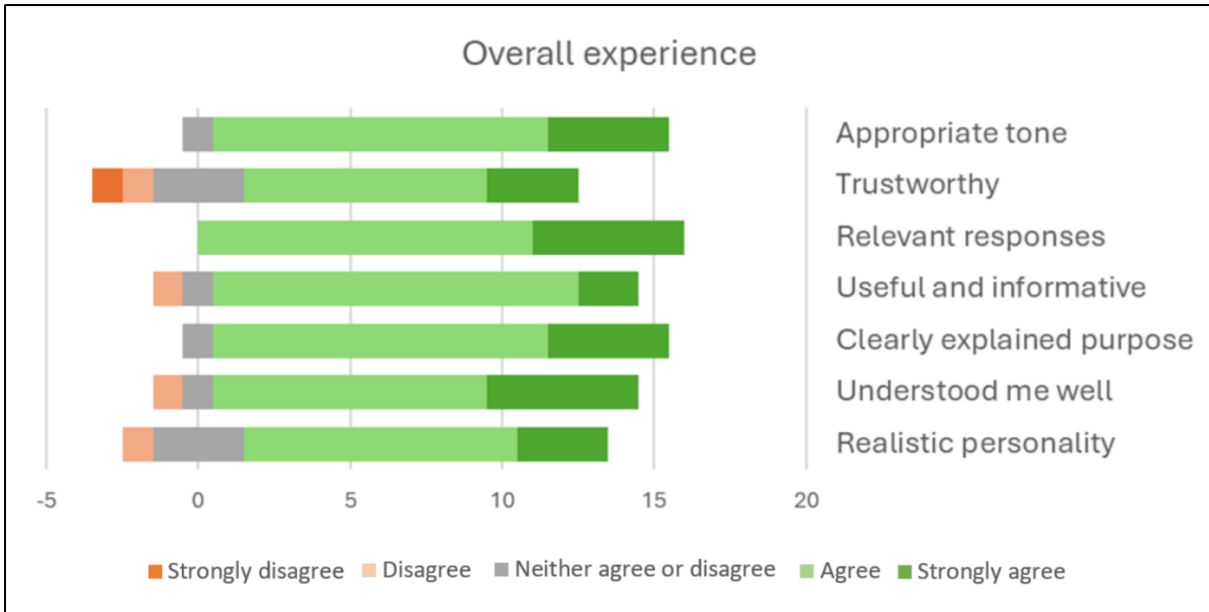


Figure 5 Participant's rating of the overall experience with the chatbot

**Quotes from the participants related to the overall experience with the chatbot:**

- *"I'm very happy with the chatbot! It needs a bit of fine tuning and then it's a great tool!"*
- *"The initial answers are quite general, but if you ask a more specific question you will get a more specific answers as well".*
- *"The chatbot responses were too slow for a normal conversation flow, I got distracted and started doing other things".*
- *"It feels like some answers were made up, which didn't feel trustworthy".*
- *"The chatbot did very well. also on uncommon questions".*
- *"A very positive experience. I got useful examples that I can use to communicate with others about TB". "*
- *"The chatbot ended each response with something like 'do you have any more questions?'. I felt that this was not so natural in the conversation".*
- *"I tried to ask several question not per se related to TB and the chatbot had a friendly way of bringing the conversation back to TB".*
- *"I asked the chatbot questions about what I can eat during my treatment and it had a friendly way of directing me to other websites. Also, I received a motivating response after I voiced that I was afraid to die from TB. Lastly, the chatbot suggested me to get tested and was very accurately able to direct me to the nearest GGD including opening hours and location".*

All participants provided a final rating for the overall experience with the chatbot. The average rating was 7.9 out of 10, ranging from 6 to 9.



The participants provided additional comments on how to improve the chatbot, such as making use of pictures to explain certain concept of TB, enable speech-to-speech interaction and clearly share the source information used.

## Costs

There are three categories for costs associated with an LLM-based chatbot. In this case, these costs include the costs of the Open Chat Studio platform, the OpenAI LLM fees and the WhatsApp fees incurred by using the WhatsApp Business Provider Twilio.

OCS monthly subscription plan: While costs may vary depending on use, in this case, the cost of Open Chat Studio was \$150 per month for the duration of the project.

OpenAI LLM costs: In this project, Dimagi used an existing API. At the start of the project, OpenAI had not yet made it possible to manage costs independently for different keys. As such, we were unable to track the costs for this round of testing. However, a project of similar size showed an average cost of \$0.33 USD per session, i.e. per single conversation for a user.

WhatsApp and web browser costs: In this particular deployment, using the chatbot on Dimagi's web browser (on Open Chat Studio) had no additional costs, as these costs are included in the OCS monthly subscription plan. The cost for the phone number used for WhatsApp conversation costs \$4.5 per month, with only \$0.80 for message related services for all conversations combined for the pilot period.

When deploying a chatbot where there is only a single interaction or conversation, the LLM costs generated would be the same whether the chatbot was deployed on WhatsApp or a web-based browser. However, when deploying a chatbot where users have multiple lengthy conversations with a bot, the LLM cost incurred on WhatsApp are higher, because the entire conversation history would be sent to the LLM with each new interaction or message sent. On a web browser, a user would start a new session each time so the past conversation history would not be sent to the LLM in the same way.

User costs: There are no cost for users when interacting with the chatbot. However, in order for users to interact with the chatbot, users should have access to a pc or mobile device with internet access.

## Discussion and conclusions

### Discussion

We have achieved all the objectives set at the start of the project to a certain extent. We assessed the **usability and acceptance** of the AI powered chatbot. Participant's feedback indicates that the chatbot is easy to use and that having a conversation with a chatbot is an accepted manner of communicating. We got good feedback on the **accuracy** of the chatbot. Overall, the accuracy has been rated highly. However, occasionally, the chatbot provided incorrect or inaccurate information. This clearly needs improvement.

In different, iterations of the chatbot, we included a variety of **safety guardrails**. In the most recent deployment of the chatbot, safety guardrails included not changing the subject away from tuberculosis in the Netherlands and specifying that the chatbot should not attempt to answer questions for which it does not have information in its source content. In one instance, after telling the chatbot they were "going to die", a user told the chatbot that they urgently needed to vomit. Given the context of the prior conversation, the chatbot assumed the user was talking about dying and vomiting in the context of tuberculosis, and responded accordingly, by emphasizing to the user that they should see a doctor as soon as possible. The chatbot also specified that the vomiting could be a result of side effects arising from TB medications but that it may also be a sign of a separate health condition.

In the set of transcripts generated, there is no example in which the chatbot provided what might be considered "harmful" or "malicious" information. Dimagi used an LLM-based tool to review both the chatbot and user responses to assess any malicious or harmful content. The tool confirmed a score of 0 on both, for each response across all transcripts, indicating that at minimum, the bot did not produce any malicious content.

We **developed a framework** to measure and evaluate the answer similarity and faithfulness of the chatbot. This framework can be used to evaluate in a standardized way, iterations of the chatbot in the future. An answer-similarity score of  $0.93 \pm 0.03$  ( $\mu \pm \sigma$ ) indicated that the answers generated were consistently on-topic. The faithfulness score of  $0.82 \pm 0.25$  ( $\mu \pm \sigma$ ) indicates that there is room for improvement. By conducting the faithfulness evaluation again after additional information is added to the source documents, we can compare the impact on the faithfulness of the chatbot. This framework can be used for every iteration of the chatbot's design and source documentation.

In the post survey and in the feedback meeting with TB professionals we made an inventory of **some benefits and limitations** of using this AI powered chatbot for improving TB knowledge for different target groups.

We determined several **barriers and facilitators for effective implementation** of an AI powered chatbot to inform patients, their families, migrant populations and the general public on TB. The main barriers are related to language skills of the chatbot and the inaccuracy of the chatbot (see more information in the section below). Facilitators are the use of the chatbot on WhatsApp, as this enables easy access. Another facilitator is the use of spoken language (speech-to-speech) to facilitate the use for people with literacy challenges.

In addition, this project has contributed considerably in **building KNCV's capacity** on the use of AI-powered chatbots. Lessons learned in this process can both contribute to the further development and implementation of AI chatbots in the Netherlands as well as in Low and Middle Income Countries (LMIC's).

While tools such as OpenAI's Assistants are meant to be very easy to use in that we can add PDFs etc as source documents, we found in this case that the existing PDFs of the educational materials could not be automatically added to the chatbot. This is because the Assistant functionality could not easily read documents with images and different types of formatting. Accordingly, we combined the available source documents in a well-formatted single source document to be used by the chatbot for reference.

## Conclusions

- We had a relatively low number of participants, therefore the participant group doesn't fully or equally represent the intended audience in terms of languages, age groups, and educational degree and prior TB knowledge.
- Overall the language skills of the chatbot are good in Dutch and English in terms of spelling, grammar and accurate communication. For less spoken languages the chatbot language skills were perceived insufficient. Language skills are part of the overall LLM and not something that can be influenced beyond the provision of source documents. As we only used English source documents for consistency purpose, perhaps using source documents in multiple languages can improve the language skills for less spoken languages.
- There is a clear correlation between language used and the overall accuracy of the answers, the chatbot was rated lower on average on all questions by those using languages beyond Dutch and English.
- We received positive feedback on the accuracy of general TB information and specific topics that were part of the source materials. Adding source information on specific topics clearly improved the chatbot's ability to provide accurate information. Providing more specific source materials on a variety of topics is likely to improve the chatbot's ability to provide detailed answers.

- Incorrect answers were provided on occasion. The margin of error should be decreased to a bare minimum if the chatbot were to be used in a programmatic setting.
- The evaluation framework for answer similarity and faithfulness scoring is a useful framework to evaluate and compare new iterations of the chatbot.
- The ease of use was rated very high, both the web browser and WhatsApp. WhatsApp is perceived as most convenient for a quick question, or to pick up from a previous conversation. The length of the answers should be a point of attention for improvement. The goal would be to strike the balance between providing an accurate response and providing too much information. Shorter answers can improve the overall flow of a conversation.
- As good as no technical challenges encountered, it is noteworthy that we piloted on a small scale. So closely monitoring the technical performance of the chatbot when used by a larger group of people over time is needed.
- The costs to operationalize the AI chatbot are manageable in a high income country. However when put to scale, the costs can become considerable higher based on the number of users, especially for LMIC's.
- We consider the pilot successful as overall the chatbot was rated a 7.9 out of 10. There are several adjustments needed before the chatbot should be used in a programmatic setting. Functionalities such as speech-to-speech optimization to lower the barrier for use of people with low literacy skills should be considered for improvement.

## Recommendations / next steps

### Stakeholder engagement

The first step is to present and share the results of this pilot project with the relevant stakeholders within the Netherlands: the RIVM, the Committee for Practical Tuberculosis Control (CPT) and representatives from the Municipal Health services (MHS) to determine if there is a sufficient support base for implementing the AI chatbot within TB care in the Netherlands. During the upcoming RIVM webinar (Oct 16 2024) on updates within Dutch TB control, the results of this project will be presented by KNCV.

### Develop an implementation plan

Developing an implementation plan should be done in close collaboration with the TB professionals and the National Institute for Public Health and the Environment (/RIVM). The implementation plan should describe the steps to improve the chatbot's design and content, and describe the roles and responsibilities of stakeholders involved, and the rollout and scale-up plan including timelines. Sustainable funding and human resource capacity for the deployment and maintenance of the AI chatbot should be taken into account in the implementation plan.

### Improvement of the chatbot

The next step to improve on the initial chatbot would be to iterate on the design and content of the chatbot. This process should include (i) the refinement of the chatbots prompt focussed on accurate, correct and shorter answers and (ii) adjustments and additions to the source documentation, including multi language source documents.

During this phase, the chatbot's improvements should be evaluated by a larger participants group that represents the target audience better in terms of languages spoken, literacy levels and educational background. The evaluation framework can be expanded upon by adding q+a pairs on additional topics added to the source documents, such as costs. And the evaluation framework can be used to evaluate and compare the improvements of the chatbot with each iteration.

### Roll-out and scale-up

After the improvements are made to the chatbot, the chatbot can be introduced to a broader audience while maintaining a focus on evaluation and continuous improvement. This phase ensures that the chatbot is effectively rolled, with careful monitoring and refinement to achieve optimal performance across diverse use cases. Throughout this stage, the chatbot's functionality, accuracy, and user engagement should be monitored closely, with iterative improvements made as needed to address any challenges that arise. Performance monitoring should be focussed on comprehensive tracking of user interaction to monitor adoption, engagement, and usage frequency.

## Applicability beyond the Dutch TB setting

This project has contributed to KNCV's and Dimagi's capacity for understanding and potential utilizing AI powered chatbots within the Dutch TB control program, and increased our understanding on the potential use of AI powered chatbots beyond the Dutch setting, both from a technical perspective and from a user-interaction perspective. Especially LMIC's, where the need for information materials and opportunities for information dissemination can be quite different than in the Dutch setting. Specifically this technology could play a role in providing relevant information on TB in the framework of supporting migrants or displaced people finding their way in the health landscape. Collaboration with International Organization for Migration (IOM) and National TB Programs (NTPs) of "recipient" countries can be considered.

