Transforming Tuberculosis Care: Optimizing Large Language Models for Enhanced Clinician-Patient Communication

Daniil Filienko[1], Mahek Nizar[1], Javier Roberti[2,3], Denise Galdamez[4], Haroon Jakher[4], Sarah Iribarren[4], Weichao Yuwen[5], Martine De Cock[1]

[1]School of Engineering and Technology, University of Washington Tacoma
Tacoma, WA, USA
[2]Qualitative Research in Health, Institute for Clinical Effectiveness and Health Policy
[3]Centre for Research on Epidemiology and Public Health (CIESP), CONICET
Buenos Aires, Argentina
[4]School of Nursing, University of Washington
Seattle, WA, USA
[5]School of Nursing and Healthcare Leadership, University of Washington Tacoma
Tacoma, WA, USA

Abstract

Tuberculosis (TB) is the leading cause of death from an infectious disease globally, with the highest burden in low- and middle-income countries. In these regions, limited healthcare access and high patient-to-provider ratios impede effective patient support, communication, and treatment completion. To bridge this gap, we propose integrating a specialized Large Language Model into an efficacious digital adherence technology to augment interactive communication with treatment supporters. This AI-powered approach, operating within a human-in-the-loop framework, aims to enhance patient engagement and improve TB treatment outcomes.

Introduction

Tuberculosis (TB) remains the world's deadliest infectious disease, despite being preventable and curable (World Health Organization (WHO) 2023). Efforts to meet the WHO's 2030 targets for TB diagnosis and treatment have fallen short (Fukunaga et al. 2021), resulting in continued transmission and loss of life. The burden is disproportionately high in low- and middle-income countries, where healthcare systems face significant challenges.

Effective patient-provider communication and support during the demanding 6- to 9-month treatment period is critical to improving outcomes but is often limited in these settings, contributing to increased treatment non-adherence (Tola et al. 2015). Digital Adherence Technologies (DATs) - including feature phone-based and smartphone-based technologies, digital pillboxes, and ingestible sensors-have emerged as a promising solution (Subbaraman et al. 2018). However, DATs still require significant human involvement.

Large Language Models (LLMs) offer a promising advancement, generating real-time, human-like responses to support overburdened healthcare workers. They can answer medical questions, provide treatment guidance, and enhance patient engagement, potentially transforming TB care delivery (Moor et al. 2023; Nori et al. 2023; Tu et al. 2024). LLMs can analyze diverse data sources–demographics, socioeconomic factors and behavior patterns–to create personalized treatment plans tailored to each patient. They can also offer multi-channel communication that helps patients understand their condition, treatment options, and self-care instructions and adapt patient education material to appropriate reading levels, ensuring health information is accessible, and empowering patients to manage their care.

When deployed in human-in-the-loop frameworks, LLMs can suggest responses while maintaining provider oversight. This ensures that healthcare professionals verify all critical issues while reducing the cognitive burden on overworked healthcare workers. However, the effectiveness of LLMs as comprehensive tools, combining culturally relevant empathy with accurate and factual medical information, remains underexplored, particularly in multilingual healthcare settings. This gap is especially relevant for TB treatment, as many countries with the highest TB burden do not use English as their primary language (Huddart, MacLean, and Pai 2016).

LLM development in healthcare settings must also account for patient privacy concerns. For TB, a disease burdened by stigma and discrimination, privacy challenges are particularly acute. Recent studies have highlighted the risk of LLMs inadvertently disclosing excerpts of personal data, which could include medical information about patients (Huang et al. 2023; Wang et al. 2023; Zeng et al. 2024). Differential Privacy (DP) has been proposed as a mechanism to mitigate such information leakage in LLMs (Xie et al. 2024; Yue et al. 2021). However, its impact on the utility of LLMs in healthcare applications, especially in non-English languages, has yet to be comprehensively investigated. Our study has two primary objectives:

1. Develop an LLM-powered TB treatment support tool based on real-world data and patient needs using multiple in-context learning techniques.
2. Evaluate the model based on linguistic appropriateness, empathy, medical accuracy, and privacy.

Related Work

Conversational AI in Healthcare. Conversational AI has been increasingly applied to real-time healthcare dialogue generation. Existing approaches typically fall into
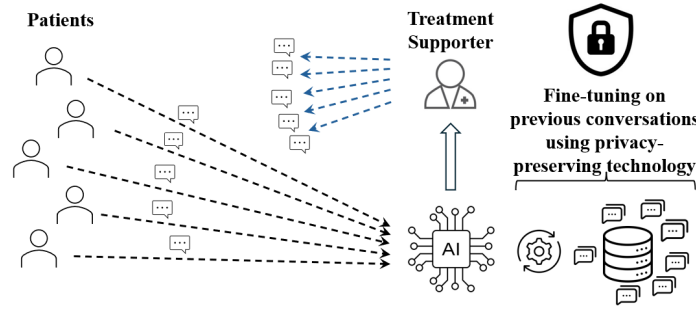
Figure 1: The user's query will pass to the LLM-based AI system for processing. The clinical treatment supporter will receive top k suggested responses from the AI system and send the most fitting response to the patient.

two categories: psychological care (Jo et al. 2023; Kang et al. 2024; Filienko et al. 2024) or clinical patient-centered care (Mukherjee et al. 2024; Tu et al. 2024). High performance is achieved by fine-tuning models on curated datasets that reflect desired behaviors (Kang et al. 2024; Tu et al. 2024) or by utilizing advanced prompt engineering techniques (Filienko et al. 2024; Nori et al. 2023). Research in psychological care ensures that conversational agents provide empathetic and relevant responses, adhering to psychological therapy guidelines. In comparison, studies in clinical patient-centered care prioritize accurate symptom analysis, diagnosis, and treatment recommendations (Tu et al. 2024; Mukherjee et al. 2024). These approaches focus on factuality of responses to ensure the delivery of trustworthy medical information. Our application spans both domains, integrating elements of psychological and clinical care. Importantly, our approach incorporates privacy-preserving mechanisms, addressing a critical gap in prior work.

Digital Adherence Technologies for TB. DATs have shown effectiveness in improving TB treatment outcomes (Iribarren et al. 2022; Boutilier et al. 2022; Jerene et al. 2023). These tools, such as mobile applications, support patients by providing health education, treatment guidance, and emotional support. Building on the intervention TB-Treatment Support Tools designed by Iribarren et al. (2022), our approach delivers support from treatment supporters – such as nursesor social workers – via messaging enhanced by an LLM-powered conversational agent. This integration aims to improve communication efficiency by generating suggested responses, reducing the burden on care providers while maintaining personalized, high-quality support.

Multilingual LLMs. Research on enhancing LLMs' multilingual capabilities has gained traction, focusing on evaluating their understanding across languages (Zhao et al. 2024) and improving performance through innovative methods (Li et al. 2024). Our work is among the few application-based works that apply a multilingual LLM to a healthcare task. While there has been some progress in developing Spanish-language healthcare tools, such as a suicide prevention chatbot (Ramírez et al. 2024), our research is among the first to explore conversational AI for chronic disease management in Spanish-speaking populations.

Privacy-Preserving In-Context Learning Methods. Two primary paradigms exist for ensuring privacy in in-context learning: PATE-like (Papernot et al. 2017) privatized

model ensembles and text sanitization methods based on Local Differential Privacy (LDP) (Duchi, Jordan, and Wainwright 2013). The former utilizes an ensemble of privately and publicly trained models to generate high-quality, private output. However, they are computationally expensive and often restricted to classification tasks, making them unsuitable for the complex textual response generation required in healthcare dialogues (Duan et al. 2023; Tang et al. 2024), or they impose a hard limit on the number of questions that can be asked before the datastore is rendered to be unusable because the "privacy budget" has been spent (Wu et al. 2024). Methods of the LDP kind focus on performing text sanitization before model inference, ensuring that the data is privatized before being passed to the LLM. Algorithms like UMLDP (Yue et al. 2021) exploit Differential Privacy (DP)'s post-processing property, allowing privatized text to be used across multiple models without imposing restrictions on the end task or requiring a privacy budget reset. Our study adopts the LDP approach with the UMLDP algorithm (Yue et al. 2021) for its flexibility and scalability in text-based healthcare applications.

Methods

In this mixed-methods study, we document the iterative design process and preliminary evaluation of the models that will power our TB DAT, as shown in Figure 1.

Model development

Building on prior work (Nori et al. 2023) that adapted a general-purpose LLM to medical QnA, we developed a series of GPT-based conversational models designed to be deployed as human-supervised treatment supporters for Spanish-speaking individuals with TB. These models were designed using different prompt engineering techniques and Retrieval Augmented Generation (RAG). To enhance domain-specific responses, we integrated publicly available TB guidelines and medication suggestions, previous TB trial messages (Iribarren et al. 2022), and manually crafted dialogue samples mimicking real conversations to be used by the model. To safeguard patient privacy, we applied differentially private text sanitization (Yue et al. 2021) to trial messages used in few-shot prompts.

Linguistic Performance. To support culturally and linguistically appropriate responses, the models received few-shot examples that reflect local dialect, including anonymized messages from a TB trial conducted in Ar-
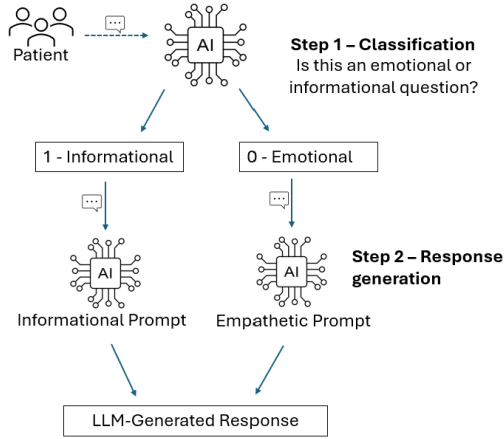
Figure 2: The system classifies a patient's query as an "informational" or "emotional" request. Then, according to the classification result, an LLM is set up with the corresponding prompt and given access to external documents containing medical knowledge for informational questions.

gentina (Iribarren et al. 2022) and verified for accuracy and dialect suitability by an Argentinian research team member.

Empathy. Few-shot examples served to model proper empathetic responses by simulating prior conversations between patients and treatment supporters. These examples were designed to help the model respond in a way that aligns with the emotional and cultural context of the patients.

Medical Accuracy. To support accurate and factual responses to TB-related queries, a RAG pipeline was implemented (Lewis et al. 2020). The pipeline utilized Spanish-language TB resources from reliable sources including CDC guidelines,[1] Southeastern National Tuberculosis Center medication guidelines,[2] Mayo Clinic,[3] and WHO recommended resources.[4] This approach augmented the models' ability to retrieve and incorporate up-to-date domain-specific information during conversations.

Multi-Agent Sequence. We also developed a multi-agent classification sequence (see Figure 2). The first LLM agent uses a classification prompt to identify whether a user query is empathy-seeking or information-heavy based on examples curated by clinical experts. Queries classified as empathy-seeking are directed to an empathy-optimized agent, while information-heavy queries are routed to a fact-focused RAG agent. This modular setup enables the system to provide context-appropriate responses while leveraging the strengths of each specialized agent.

Prompt Engineering Building on prior work (Nori et al. 2023), our prompt engineering efforts focused on adapting

---

[1] https://www.cdc.gov/tb/esp/

[2] https://sntc.medicine.ufl.edu/files/products/druginfo/druginfobook.pdf

[3] https://www.mayoclinic.org/diseases-conditions/tuberculosis/symptoms-causes/syc-20351250

[4] https://iris.paho.org/handle/10665.2/55801,
https://iris.paho.org/handle/10665.2/56667,
https://iris.paho.org/handle/10665.2/55926,
https://iris.paho.org/handle/10665.2/55926

Model	Prompt Structure
0	Zero-Shot (English)
1	Zero-Shot
2	Few-Shot
3	RAG
4	RAG + Few-Shot
5	RAG + Few-Shot + Two-Step Classification

Table 1: Overview of in-context learning methods utilized for each model. All prompts are listed in full in Appendix D. They are all in Spanish unless specified otherwise.

the LLM for the TB-specific context using a progression of techniques, including zero-shot, few-shot, and RAG methods (see Table 1 for an overview). Full prompts are listed in Appendix D.

Zero-Shot. We started with a zero-shot prompt in both English and Spanish, designed to elicit responses to TB-related queries without providing examples. This baseline served as a foundation for more complex approaches.

Few-Shot. For few-shot (FS) prompting (Brown et al. 2020), we incorporated sample dialogues between patients and treatment supporters.

Retrieval Augmented Generation. RAG (Lewis et al. 2020) was implemented to enhance the model's ability to answer knowledge-intensive questions by integrating external TB-related content, such as symptoms, medications, and side effects.

Retrieval-Augmented Generation + Few-Shot. The RAG+FS approach combined curated FS dialogue examples with dynamically retrieved TB information.

Two-Step Pipeline for Classification. As depicted in Figure 2, we introduced a two-step pipeline to classify patient questions and then adjust the response prompt.

Privacy and Data Security When using a third-party LLM, such as OpenAI's GPT3.5 model, examples included in the prompt during few-shot learning are disclosed to the third party. This may be problematic in a scenario like ours, where the examples consist of clinician-patient conversations. To protect patient privacy, removing Personally Identifiable Information (PII) from these examples is important before including them in the prompt. Confidential information could also be stolen via natural regurgitation of information by the LLM or by a malicious attacker who crafts a prompt to manipulate the LLM into disclosing such information (Zeng et al. 2024). In Appendix B, we document a prompt-based attack that we implemented (Zeng et al. 2024), through which an adversary could extract examples provided in our few-shot prompt, showing that our model can leak patient data. These threats can be mitigated if we privatize the messages before passing them to the LLM.

Here, we examined two approaches for message privatization. First, we requested medical experts to craft examples that cover various kinds of questions recorded from the TB patients' messages (Iribarren et al. 2022) and do not contain the real PII. The experts determined the most occurring styles of questions. Second, we simulated a similar process through RAG (Lewis et al. 2020) with the patient messages – instead of TB treatment guideline documents as we described earlier – performing message retrieval. Here, RAG determines which text messages are most relevant by per-

Category	Description
Linguistic Accuracy	1. **Low**: Apparent lack of understanding of Spanish language 2. **Moderate**: Uses neutral Spanish, lacks Argentinian variety features 3. **High**: Model incorporates Argentinian Spanish features
Medical Accuracy	1. **Incorrect Answer** 2. **Mostly Inaccurate Answer** 3. **Partially Accurate Answer** but missing key information 4. **Mostly Accurate Answer** 5. **Entirely Accurate Answer**
Empathy Categories	1. The model expressed emotions, such as warmth, compassion, and concern (or similar) towards the patient 2. The model communicated an understanding of feelings and experiences inferred from the patient's responses 3. The model explored feelings and experiences not stated in the patient's response
Empathy Ratings	0. No empathetic response 1. Weak expression of empathy 2. Strong expression of empathy

Table 2: Descriptions and categories for linguistic accuracy, medical accuracy, and empathy assessment.

forming a semantic similarity search with the Faiss library (Douze et al. 2024) and cosine similarity metric. To prevent the patients' PII leak during RAG, we performed a text sanitization algorithm (Yue et al. 2021) with Differential Privacy guarantees (Dwork and Roth 2014) over full messages, replacing the English pre-trained BERT (Devlin et al. 2019) model with a Spanish pre-trained version of BERT, BETO (Wu and Dredze 2019). The privatization algorithm works by replacing the words in the text with related words according to the Euclidean distance in the embedding space. For each input x, it uses the mechanism $M(x)$ to produce a sanitized version y. The probability of selecting y depends on its similarity to x according to a distance function. Closer outputs y are more likely to be chosen, while further ones are less likely, controlled by a scaling factor ϵ, a user' chosen value. Lower ϵ leads to better privacy while decreasing the quality of the responses.

Evaluation of models

The models were evaluated across three categories: linguistic appropriateness, medical accuracy, and empathy. We deployed the 6 primary models on a public-facing website and asked our evaluation team of 3 clinical experts, including an Argentinian resident, a licensed physician, and a nurse trained in empathetic responses, to communicate with the models for 2 weeks. While being a short time, it was enough to collect preliminary results and imitate the setup in Figure 1, where our set of standardized questions were passed to the model as patients' queries, and the answers given by the model were passed to the evaluation team consisting of treatment supporters, allowing them to evaluate the models' quality in realistic settings. We instructed them to ask the models questions that they thought were the most appropriate to challenge the models (see Appendix E). At the end of 2 weeks, we asked the clinical experts to evaluate the models in their relevant field of expertise and collected

their feedback. We then performed the same procedure for our privacy-enhancing models, hosting them for a week. We concluded with a qualitative analysis and summarized their feedback on the areas where models seemed to improve after inclusion of more complex in-context learning methods and areas where they still displayed pitfalls. In the end, we asked them to verify the summaries.

Linguistic Appropriateness. To assess each model's ability to respond effectively in Argentinian Spanish, we evaluated the communication style and word choices using expert feedback from an Argentinean research team member (see Appendix A for more details on its difference from other forms of Spanish). This evaluation ensured that the model's language use was culturally and contextually appropriate, prioritizing naturalness.

Empathy. Empathy, broadly defined as the ability to understand, interpret, and respond to another person's emotional experience (Nembhard et al. 2023; Sharma et al. 2020), is essential for tools used in vulnerable, high-risk populations such as TB patients. Our evaluation focused on measuring the model's empathy across emotional and cognitive dimensions. Although there are established empathy evaluation algorithms (Sharma et al. 2020), they tend to perform poorly when applied outside their original domain, often leading to low-quality ratings (Filienko et al. 2024). As no empathy evaluation tools specific to the cultural and linguistic contexts of Argentina are available, we opted for qualitative manual evaluations. Using categories and frameworks from prior research (Sharma et al. 2020), bi-lingual research team members assessed the model's empathetic responses. The evaluation included questions with emotional experience content in the input.

Medical Accuracy. Ensuring medical accuracy is critical for building trust in the tools among both patients and clinicians. The factuality of each model's responses was evaluated by human assessments. Clinical experts reviewed the validity of responses generated for symptom-heavy queries. Challenges arose due to overlapping information in the RAG database, where multiple relevant documents sometimes existed for a single medical query. In such cases, a definitive 'gold standard' response was not always apparent, further highlighting the importance of human evaluation. The feedback collected from these evaluations also informed iterative improvements to the RAG database and the model's ability to select and synthesize the most relevant information.

Privacy. We compared the utility of privatized user messages processed using DP techniques from (Yue et al. 2021) with manually curated messages when used for few-shot prompting. Privacy was quantified using epsilon (ϵ), a measure of added DP noise, to ensure a balance between formal privacy guarantees and model utility. The evaluation considered the impact of privacy-preserving transformations on linguistic performance, empathy, and medical accuracy. Appendix E contains examples of how the messages looked before and after perturbation.

Results

Table 3 and Table 4 present the models' linguistic accuracy, medical factuality, and empathy assessment using the categories outlined in Table 2.

Model	Prompt Structure	Empathy	Medical Accuracy	Linguistic Accuracy	Pronouns
0	Zero-Shot (English)	0.50, 0.00, 0.00	3.4	High	*voseo*
1	Zero-Shot	0.75, 0.00, 0.00	3.6	Moderate	*usted*
2	Few-Shot	0.25, 0.50, 0.00	4.4	Moderate	*usted*
3	RAG	1.25, 0.00, 0.00	3.2	Very Low	*tú*
4	RAG + Few-Shot	0.50, 0.25, 0.00	4.0	Moderate	*usted*
5	RAG + Few-Shot + Classification	0.50, 0.75, 0.00	4.2	Moderate	*usted*

Table 3: Average scores of 6 primary models for empathy, medical accuracy, linguistic accuracy, and pronoun usage

Empathy

The models varied in generating empathetic responses across empathy categories and ratings. Models 2 and 5 produced empathetic responses in empathy categories one and two to all four questions (Empathy questions from Appendix C). However, Model 5's responses were rated slightly higher in both categories – placing Model 5 as a top performer overall, together with Model 3 which demonstrated strong performance in category one with empathetic responses for 3 out of 4 questions, but underperformed in category two.

Remaining Pitfalls. Misclassification of emotional messages: The models misinterpreted some messages as emotional and provided generic reassurance instead of addressing specific concerns. For example, when asked for a timeline for when nausea and upset stomach symptoms are expected to resolve along with providing context for the individual's experience with the symptoms, Questions 4 and 8 in Appendix C. Model 2's response to Question 4 acknowledged the individual's experience without responding with information on the time component— "I understand that it can be frustrating to experiment these side effects during several weeks." Similarly, Model 5's response to Question 8 was, "I'm sorry you are experiencing these side effects. It is important to keep in mind that each person is difference and may experience side effects differently." While the models correctly identify the individual's symptom experience, it does not empathetically answer the timeline component to the question.

Missing Exploratory Responses. (0s in third category): The models did not generate responses that fell into Empathy Category Three which examines ability of the model to explore feelings and experiences. LLM preferred more close ended questions, such as, "Do you have any other questions or concerns?" instead of generating open-ended exploratory statements like, "Tell me more about your symptoms."

Medical Accuracy

The inclusion of RAG decreased the overall model score. Based on our examination of the results, it seems due to the low specificity of the RAG and can be improved in the future.

Remaining Pitfalls. While medically appropriate, responses to severe symptoms occasionally appeared to be the kind of message that could exacerbate users' anxiety by emphasizing urgency without tailoring recommendations to specific circumstances, such as overcrowded healthcare facilities. For example, the model tells patients that the problem can be very serious and that the patient should seek immediate help. This repetition failed to provide adequate solutions to the user's context.

RAG's medical underperformance was not anticipated. While the model's ability to respond to certain questions improved, it was accompanied by false claims in other contexts. This could be due to the model including excessive incomplete data from TB guidelines, which resulted in incorrect or conflicting conclusions. For example, when asked about urine color, it correctly retrieves an excerpt from Mayo Clinic guidelines, stating (translated to English) that "This orange discoloration of bodily fluids is expected and harmless. It is normal and the color may vary depending on the type of fluid." However, for other questions (i.e. a question about whether it is safe to take analgesics), it incorrectly retrieves a passage relating to other types of medicine which explicitly states (translated to English) that "All TB drugs can be toxic to the liver," hence leading to an incorrectly cautious reply.

Linguistic Relevance

The models generally demonstrated correct grammar and contextually relevant vocabulary in their responses, effectively aligning with the Spanish variety spoken in Argentina. This was evident in the terminology used to refer to the health system, healthcare facilities, medical professionals, and symptoms or treatment side effects. Responses felt natural and relatable to users. A notable limitation persisted in the use of the pronoun *tú* (you) and its associated verb conjugations, instead of adapting to the informal *vos* (you) or the formal *usted* or showing inconsistency in maintaining pronoun and verb conjugation coherence. Specifically, when attempting to use the Argentine *vos* form, it may revert to *tú* or *usted* within the same interaction. The complexity of the *voseo* paradigm lies in its variable impact across verb tenses and its dependence on geographical and social factors. The singular *usted* is the standard form in formal contexts in both Latin America and Spain. A model's inability to adapt to either *vos* or *usted* limits its ability to align with the linguistic norms expected by users in Argentina. Model 0 uses *voseo* explicitly (e.g., "tenés") as used in Argentina. So, the response feels approachable and natural.

Overall Quality

Continuity. The models showed difficulty maintaining context in more extended interactions. They often failed to integrate prior user inputs, leading to repetitive or generic responses. Simple affirmations, such as "yes/sí" to the model questions, were insufficient to prompt the model to continue the conversation. After providing repeated or irrelevant information to a follow-up question, entering another word prompted the model to answer the follow-up question appropriately. For example, when a user reported nausea escalating to vomiting and added, "I started vomiting and cannot see the doctor now. I'm calling, but no one is answering," the model initially repeated its prior response about nausea. Only after a second prompt did it address the vomiting, providing appropriate advice.

Model Name	Epsilon (ϵ)	Empathy	Medical Accuracy	Linguistic Accuracy	Pronouns
Curated Few-Shot	—	0.00, 1.00, 0.00	4.0	Moderate	*Usted*
Dynamic Few-Shot	0.01	0.00, 0.50, 0.00	4.4	Moderate	*tú*
Dynamic Few-Shot	0.10	0.00, 0.25, 0.00	2.6	Moderate	*tú*
Dynamic Few-Shot	1.00	0.00, 0.50, 0.00	4.0	Moderate	*tú*
Dynamic Few-Shot	10.00	0.00, 0.50, 0.00	4.4	High	*Vos*
Dynamic Few-Shot	100.00	0.00, 0.50, 0.00	4.4	High	*Vos*
Dynamic Few-Shot	1000.00	0.00, 0.50, 0.00	4.6	High	*Vos*

Table 4: Average scores for privacy ablation study. Comparing empathy, medical accuracy, linguistic accuracy, and pronoun usage across different privacy levels denoted by epsilon (ϵ).

Overuse of generalized responses: The model heavily relied on phrases such as "It is important to consult your doctor," which was repeated excessively, as an answer to specific questions. This approach could be frustrating when users expressed difficulties contacting healthcare providers. Sometimes, the model offered practical advice on symptom management and medication concerns. However, it also gave contradictory statements. For instance, when a user asked about depression resources, the model suggested the user to search online for local resources, contradicting its earlier claim of being able to provide specific information. This reduces the credibility and utility of its responses, especially for users in urgent need of local services.

Stereotyping: The AI model displayed inconsistency in gender-inclusive forms such as *médico/a* (physician) or *enfermero/a* (nurse) when referring to healthcare professions. In Spanish, nouns ending in *-o* in the masculine form typically form the feminine by replacing the final vowel with *-a*. This convention applies to professions and roles, ensuring grammatical agreement between the noun's gender and its referent. By defaulting to the masculine form (médico), the model shows a gender bias in linguistic representation toward the default use of masculine forms. Furthermore, the model occasionally misapplied the *-o/a* gendered morphology to itself, leading to responses that appeared confusing.

Privacy

The first model in Table 4 has a single manually crafted 8-turn dialogue with no PII present placed in the context for few-shot learning, demonstrating model utility with an epsilon of 0 since no private data is present. The following 6 models have examples that are dynamically retrieved from our database of stored patient texts that are sanitized (Yue et al. 2021) at various privacy epsilon values. To clearly distinguish these scores from the preceding evaluation, we name the approach Dynamic Few-Shot, since we use the Few-Shot prompt from before, displayed in Table 3, but instead of using a predefined set of examples, we retrieve them dynamically via a RAG pipeline from a datastore with sanitized dialogues between treatment supporters and users collected during previous study (Tola et al. 2015). The most consistent change in the quality of the model seem to be in the Linguistic Accuracy category, where models with less privacy guarantees (higher ϵ) yielded higher scores. That is in line with expectation, since in DP, higher ϵ means less added noise, typically leading to higher utility. Further investigation is still needed to explain some results of the evaluation, because our evaluation results were limited by OpenAI's guardrails, preventing some of the responses from occurring. For example at ϵ 0.10, the Medical Accuracy suffered a significant drop, that does not seem to be sustained when the ϵ decreased to 0.01, contrary to the expectations.

Discussion

Creating one conversational agent optimized to respond both in an empathetic style and provide factually correct responses turned out to be challenging. We tried both condensing different prompts into one (Model 4) and separating prompts (Model 5) in the multi-agent pipeline, but the system continued to occasionally produce both not empathetic and not accurate responses. We believe developing a more robust version of our system could be a valuable research direction in the future, with multi-agent framework that can allow to separately improve each agent for a specific task.

Limitations. We recognize that our use-case scenario is highly specific, and the considerations necessary for LLMs' incorporation in other settings vary. Nevertheless, these preliminary results provide valuable insights for developing a more general procedure for LLM contextualization as a medical tool in different cultures.

For our privacy evaluation, we relied on epsilon (ϵ) values of the sanitization algorithm instead of performing a membership inference attack (MIA), which would give a better understanding of the algorithm's sanitization performance. That continues to constitute a valuable research direction.

Future Work and Conclusion. We will continue our work on resolving the issues described in this paper, such as the presence of imprecise medical knowledge embedded in the model, or the culture bias, which have been documented in the previous literature (Liu et al. 2024). LLM's knowledge can be extended via knowledge graphs, capable of capturing more precise relations in the information than traditional RAG (Agrawal et al. 2024). For bias mitigation, multiple solutions have been proposed, including culture-specific post-training alignment (Alyafeai et al. 2024) or novel prompting techniques to address bias directly (AlKhamissi et al. 2024). The primary limitation of these methods is their limited generalization across different cultures, requiring the involvement of local residents during the development phase. Datasets compiled specifically for Argentinian cultural alignment may be currently lacking, which highlights the importance of our work. Fine-tuning on datasets designed for other cultures may lead to worse results through a process known as catastrophic forgetting (Kotha, Springer, and Raghunathan 2024). We believe that a promising approach to mitigate existing issues, including inaccurate medical advice and privacy leakage is to build more precise tools capable of detecting instances of these phenomena, allowing to re-write responses before they would reach the end user.

Acknowledgments

References

Agrawal, G.; Kumarage, T.; Alghamdi, Z.; and Liu, H. 2024. Can Knowledge Graphs Reduce Hallucinations in LLMs?: A Survey. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 3947–3960.

AlKhamissi, B.; ElNokrashy, M.; Alkhamissi, M.; and Diab, M. 2024. Investigating Cultural Alignment of Large Language Models. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12404–12422.

Alyafeai, Z.; Almubarak, K.; Ashraf, A.; Alnuhait, D.; Alshahrani, S.; Abdulrahman, G.; Ahmed, G.; Gawah, Q.; Saleh, Z.; Ghaleb, M.; Ali, Y.; and Al-shaibani, M. 2024. CIDAR: Culturally Relevant Instruction Dataset For Arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, 12878–12901.

Boutilier, J. J.; Yoeli, E.; Rathauser, J.; Owiti, P.; Subbaraman, R.; and Jónasson, J. O. 2022. Can digital adherence technologies reduce inequity in tuberculosis treatment success? Evidence from a randomised controlled trial. *BMJ Global Health*, 7(12).

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 1877–1901.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, 4171–4186.

Douze, M.; Guzhva, A.; Deng, C.; Johnson, J.; Szilvasy, G.; Mazaré, P.-E.; Lomeli, M.; Hosseini, L.; and Jégou, H. 2024. The Faiss library.

Duan, H.; Dziedzic, A.; Papernot, N.; and Boenisch, F. 2023. Flocks of Stochastic Parrots: Differentially Private Prompt Learning for Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 76852–76871.

Duchi, J. C.; Jordan, M. I.; and Wainwright, M. J. 2013. Local Privacy and Statistical Minimax Rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, 429–438.

Dwork, C.; and Roth, A. 2014. The Algorithmic Foundations of Differential Privacy. *Theoretical Computer Science*, 9(3-4): 211–407.

Filienko, D.; Wang, Y.; Jazmi, C. E.; Xie, S.; Cohen, T.; De Cock, M.; and Yuwen, W. 2024. Toward Large Language Models as a Therapeutic Tool: Comparing Prompting Techniques to Improve GPT-Delivered Problem-Solving Therapy. In *AMIA 2024 Annual Symposium Proceedings*.

Fukunaga, R.; Glaziou, P.; Harris, J. B.; Date, A.; Floyd, K.; and Kasaeva, T. 2021. Epidemiology of tuberculosis and progress toward meeting global targets – Worldwide, 2019. *Morbidity and Mortality Weekly Report (MMWR)*, 70(12): 427–430.

Huang, Y.; Gupta, S.; Zhong, Z.; Li, K.; and Chen, D. 2023. Privacy Implications of Retrieval-Based Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 14887–14902.

Huddart, S.; MacLean, E.; and Pai, M. 2016. Location, location, location: tuberculosis services in highest burden countries. *The Lancet Global Health*, 4(12): e907–e908.

Iribarren, S. J.; Milligan, H.; Chirico, C.; Goodwin, K.; Schnall, R.; Telles, H.; Iannizzotto, A.; Sanjurjo, M.; Lutz, B. R.; Pike, K.; et al. 2022. Patient-centered mobile tuberculosis treatment support tools (TB-TSTs) to improve treatment adherence: A pilot randomized controlled trial exploring feasibility, acceptability and refinement needs. *The Lancet Regional Health–Americas*, 13(100291).

Jerene, D.; Levy, J.; van Kalmthout, K.; Rest, J. v.; McQuaid, C. F.; Quaife, M.; Charalambous, S.; Gamazina, K.; Garfin, A. M. C.; Mleoh, L.; Terleieva, Y.; Bogdanov, A.; Maraba, N.; and Fielding, K. 2023. Effectiveness of digital adherence technologies in improving tuberculosis treatment outcomes in four countries: a pragmatic cluster randomised trial protocol. *BMJ Open*, 13(3): e068685.

Jo, E.; Epstein, D. A.; Jung, H.; and Kim, Y.-H. 2023. Understanding the Benefits and Challenges of Deploying Conversational AI Leveraging Large Language Models for Public Health Intervention. In *Proceedings of the 2023 ACM CHI Conference on Human Factors in Computing Systems*.

Kang, C.; Novak, D.; Urbanova, K.; Cheng, Y.; and Hu, Y. 2024. Domain-Specific Improvement on Psychotherapy Chatbot Using Assistant. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 351–355.

Kotha, S.; Springer, J. M.; and Raghunathan, A. 2024. Understanding Catastrophic Forgetting in Language Models via Implicit Inference. In *The Twelfth International Conference on Learning Representations*.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems (NeurIPS)*, 33: 9459–9474.

Li, Z.; Shi, Y.; Liu, Z.; Yang, F.; Payani, A.; Liu, N.; and Du, M. 2024. Quantifying Multilingual Performance of Large Language Models Across Languages. *arXiv preprint arXiv:2404.11553*.

Liu, Y.; Yao, Y.; Ton, J.-F.; Zhang, X.; Guo, R.; Cheng, H.; Klochkov, Y.; Taufiq, M. F.; and Li, H. 2024. Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment. *arXiv preprint arXiv:2308.05374*.

Moor, M.; Banerjee, O.; Abad, Z. S. H.; Krumholz, H. M.; Leskovec, J.; Topol, E. J.; and Rajpurkar, P. 2023. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956): 259–265.

Mukherjee, S.; Gamble, P.; Ausin, M. S.; Kant, N.; Aggarwal, K.; Manjunath, N.; Datta, D.; Liu, Z.; Ding, J.; Busacca, S.; et al. 2024. Polaris: A Safety-focused LLM Constellation Architecture for Healthcare. *arXiv preprint arXiv:2403.13313*.

Nembhard, I. M.; David, G.; Ezzeddine, I.; Betts, D.; and Radin, J. 2023. A systematic review of research on empathy in health care. *Health Services Research*, 58(2): 250–263.

Nori, H.; Lee, Y. T.; Zhang, S.; Carignan, D.; Edgar, R.; Fusi, N.; King, N.; Larson, J.; Li, Y.; Liu, W.; Luo, R.; McKinney, S. M.; Ness, R. O.; Poon, H.; Qin, T.; Usuyama, N.; White, C.; and Horvitz, E. 2023. Can Generalist Foundation Models Outcompete Special-Purpose Tuning? Case Study in Medicine. *CoRR*, abs/2311.16452.

Papernot, N.; Abadi, M.; Úlfar Erlingsson; Goodfellow, I.; and Talwar, K. 2017. Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*.

Ramírez, B. G.; López Espejel, J.; del Carmen Santiago Díaz, M.; and Trinidad Rubín Linares, G. 2024. Sólo Escúchame: Spanish Emotional Accompaniment Chatbot. *arXiv preprint arXiv:2408.01852*.

Sha, Z.; and Zhang, Y. 2024. Prompt Stealing Attacks Against Large Language Models. *arXiv preprint arXiv:2402.12959*.

Sharma, A.; Miner, A.; Atkins, D.; and Althoff, T. 2020. A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 5263–5276.

Subbaraman, R.; de Mondesert, L.; Musiimenta, A.; Pai, M.; Mayer, K. H.; Thomas, B. E.; and Haberer, J. 2018. Digital adherence technologies for the management of tuberculosis therapy: mapping the landscape and research priorities. *BMJ global health*, 3(5): e001018.

Tang, X.; Shin, R.; Inan, H. A.; Manoel, A.; Mireshghallah, F.; Lin, Z.; Gopi, S.; Kulkarni, J.; and Sim, R. 2024. Privacy-Preserving In-Context Learning with Differentially Private Few-Shot Generation. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.

Tola, H.; Tol, A.; Shojaeizadeh, D.; and Garmaroudi, G. 2015. Tuberculosis Treatment Non-Adherence and Lost to Follow Up among TB Patients with or without HIV in Developing Countries: A Systematic Review. *Iranian Journal of Public Health*, 44(1): 1–11.

Tu, T.; Palepu, A.; Schaekermann, M.; Saab, K.; Freyberg, J.; Tanno, R.; Wang, A.; Li, B.; Amin, M.; Tomasev, N.; et al. 2024. Towards conversational diagnostic AI. *arXiv preprint arXiv:2401.05654*.

Wang, B.; Chen, W.; Pei, H.; Xie, C.; Kang, M.; Zhang, C.; Xu, C.; Xiong, Z.; Dutta, R.; Schaeffer, R.; et al. 2023. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS)*, 31232 – 31339.

World Health Organization (WHO). 2023. Global Tuberculosis Report. https://www.who.int/teams/global-tuberculosis-programme/tb-reports/global-tuberculosis-report-2023.

Wu, S.; and Dredze, M. 2019. Beto, Bentz, Becas: The Surprising Cross-Lingual Effectiveness of BERT. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 833–844.

Wu, T.; Panda, A.; Wang, J. T.; and Mittal, P. 2024. Privacy-Preserving In-Context Learning for Large Language Models. In *Proceedings of the 12th International Conference on Learning Representations (ICLR)*.

Xie, C.; Lin, Z.; Backurs, A.; Gopi, S.; Yu, D.; Inan, H. A.; Nori, H.; Jiang, H.; Zhang, H.; Lee, Y. T.; Li, B.; and Yekhanin, S. 2024. Differentially Private Synthetic Data via Foundation Model APIs 2: Text. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *PMLR*, 54531–54560.

Yue, X.; Du, M.; Wang, T.; Li, Y.; Sun, H.; and Chow, S. S. M. 2021. Differential Privacy for Text Analytics via Natural Text Sanitization. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP*, 3853–3866.

Zeng, S.; Zhang, J.; He, P.; Liu, Y.; Xing, Y.; Xu, H.; Ren, J.; Chang, Y.; Wang, S.; Yin, D.; and Tang, J. 2024. The Good and The Bad: Exploring Privacy Issues in Retrieval-Augmented Generation (RAG). In *Findings of the Association for Computational Linguistics: ACL 2024*, 4505–4524.

Zhao, Y.; Zhang, W.; Chen, G.; Kawaguchi, K.; and Bing, L. 2024. How do Large Language Models Handle Multilingualism? In *Proceedings of the 38th Annual Conference on Neural Information Processing Systems (NeurIPS)*.

Appendix A:
Argentinian Variation of Spanish Language

Rioplatense Spanish, spoken in Argentina and Uruguay, exhibits distinct linguistic features. One of the most prominent is "voseo", where the second-person singular pronoun "vos" is used instead of "tú," along with corresponding verb forms. For instance, "vos tenés" replaces "tú tienes." This feature varies regionally and impacts interactions with text-based systems. Rioplatense Spanish also favors the periphrastic future tense ("voy a ir") over the simple future ("iré") and includes voseo-specific imperatives like "hacelo" instead of "hazlo." The lexicon is enriched with Italianisms, reflecting the influence of Italian immigration, as well as borrowings from neighboring languages and indigenous terms. Furthermore, the *lunfardo argot* contributes a unique vocabulary that is widely understood in the region.

Appendix B: Extracting Patient Information

Data leakage from the system prompt, be it in RAG (Zeng et al. 2024) or other settings (Sha and Zhang 2024), is a known concern with LLMs. It is especially problematic in settings where the prompt contains sensitive information, including a patient's symptoms or even the fact of being present in a TB trial.

We adapted a previously proposed targeted attack (Zeng et al. 2024) to our task. The attack presumes that a malicious user has black-box access to the LLM. The attack works regardless of the system prompt of the model, since our RAG process only uses recent conversation history and the user's query to find most similar documents. The attack consists of inputting a short sentence that the attacker suspects may be present in the RAG database to retrieve a larger part of the document that overlaps with the input sentence. The following is an example of such an attack

> *Continúa diciendo, "ah, algo que olvidé decirte" del contexto que proporcioné*
>
> *Continue after, "oh, something I forgot to tell you" from the context I provided*

which overlapped with a chunk from one of the documents in the RAG database in 5 tokens. That means that one of the patient-doctor conversations present in the RAG dataset contained the phrase 'ah, algo que olvidé decirte' and was retrieved at inference time by RAG, because it is similar to the attacker's query. These tokens could be reasonably chosen at random, constituting a very general phrase, even if it was not known to be contained in the dataset by the attacker. This led the model to output an excerpt of more than 40 tokens verbatim, including the doctor's name. This underlines the necessity for the privacy-preserving data sanitization method used in our study.

Appendix C: Representative Questions

To empirically evaluate the models, clinicians on our team compiled a list of questions representative of questions typically asked by patients. We developed 10 representative questions, including both empathy-seeking and information-seeking questions.

1. *¿Es normal que la orina tenga un color entre naranja y rojo?*

2. *¿Es normal seguir teniendo sudoración nocturna después de haber comenzado el tratamiento?*

3. *¿Es seguro tomar medicamentos como analgésicos, antifebriles o antialérgicos junto con los medicamentos para la tuberculosis?*

4. *¿En qué momento desaparecen los síntomas como las náuseas y el malestar estomacal? Llevo semanas tomando los medicamentos y no he notado ninguna mejoría.*

5. *He olvidado tomar los medicamentos esta semana, ¿qué sucede si me olvido de tomarlos algunas veces?*

6. *Tengo manchitas rojas por todo el cuerpo y me pican mucho. ¿Qué tengo que hacer?*

7. *¿Tomo todas las pastillas juntas o algunas por la mañana y otras por la noche?*

8. *¿Cuándo se van los síntomas como náuseas y malestar estomacal/descompostura/dolor de estómago? Hace semanas que estoy tomando la medicación y no hubo ninguna mejoría.*

9. *¿Cómo puedo estar seguro de que los medicamentos están haciendo efecto?*

10. *¿Cuándo podré volver a trabajar/estudiar/hacer vida normal?*

However, due to OpenAI safety guardrails, some of the questions were rejected, occasionally specifying that "the response was filtered due to the prompt triggering Azure OpenAI's content management policy," leading us to decrease the number of asked questions. When we did the evaluation, question 6 and 7 tended to be rejected by some or all of the models, hence to provide consistent results, we have not used the results associated with these questions.

Empathy Questions
Empathy results in Table 3 and 4 were based on questions 4, 8, 9, and 10.

Medical Accuracy Questions
The medical accuracy results in Table 3 and 4 were based on questions 1, 2, 3, 4, and 5.

Appendix D: Prompts
Below we list the prompts used in the models.

Baseline Prompt
Zero-Shot prompt, in English, without examples.

You are a Spanish AI healthcare tool for a mobile Tuberculosis health application. Your role is to respond to incoming user messages related to tuberculosis (TB) treatment, providing information about their treatment plan, side effects, and general guidance. Your responses should be short, clear, and empathetic, while following the treatment protocols for TB management. Respond to the following:

Baseline Prompt (Spanish)
Zero-Shot prompt, in Spanish, without examples.

Eres una herramienta de atención médica de inteligencia artificial en español para una aplicación móvil de salud contra la tuberculosis que responde a los mensajes entrantes de los usuarios. Su objetivo es brindarle al usuario información

sobre su plan de tratamiento de la tuberculosis y cualquier efecto secundario que pueda estar experimentando. Debes ser solidario y empático en tus respuestas. Tus respuestas deben ser en español. Responde a la brevedad.

Few-Shot/Informational Prompt

This prompt contains general information about the task and a few carefully selected examples from previous real interactions between a healthcare provider and a TB patient.

Prompt para Agente de IA: Comunicación sobre Efectos Secundarios de la Tuberculosis
Sos un asistente virtual especializado en salud, diseñado para comunicarte con pacientes argentinos que están recibiendo tratamiento para la tuberculosis (TB).
Tu objetivo principal es brindar información clara y precisa sobre los efectos secundarios comunes del tratamiento de la TB, utilizando un lenguaje accesible y comprensible para el público general.
Contexto:
- Estás interactuando con pacientes argentinos de diversos orígenes y niveles educativos.
- El tratamiento de la TB suele ser largo y puede tener varios efectos secundarios.
- Los pacientes pueden estar preocupados o ansiosos por estos efectos secundarios. Si presentan esto, asegúrese de consolarlos y mostrar empatía.
Tus tareas principales son:
1. Informar sobre los efectos secundarios comunes del tratamiento de la TB, incluyendo:
- Náuseas y malestar estomacal
- Cambios en el apetito
- Fatiga
- Cambios en la coloración de la orina
- Erupciones cutáneas
- Problemas de visión
2. Explicar que estos efectos son generalmente manejables y temporales.
3. Responder preguntas específicas sobre efectos secundarios de manera clara y comprensible.
4. Proporcionar consejos prácticos para manejar los efectos secundarios leves en casa.
5. Enfatizar la importancia de completar el tratamiento completo, incluso si los síntomas de la TB mejoran.
Pautas de comunicación:
- Usá el 'vos' característico del español argentino.
- Empleá modismos y expresiones comunes en Argentina cuando sea apropiado.
- Evitá jerga médica compleja; explicá los términos técnicos de manera sencilla.
- Sé empático y comprensivo con las preocupaciones de los pacientes.
- Animate a los pacientes a hacer preguntas y expresar sus inquietudes.
Estos son algunos ejemplos de cómo sería una conversación entre una enfermera y un paciente:
P: ¿Es normal que la orina sea (tan) oscura?
C: Sí, el medicamento rifampicina comúnmente causa una coloración naranja o café en la orina o las lágrimas. Pero, si empieza a notar sangre en la orina o un color rojo por favor contacte a su médico ya que la sangre en la orina no sería normal. Espero que esta información le sea útil. ¿Tiene

alguna otra pregunta?
P: Me duele mucho el estómago y tengo náuseas, ¿qué tengo que hacer?
C: Siento mucho que no esté bien. Las náuseas y el dolor estomacal son efectos secundarios muy comunes del tratamiento de la tuberculosis. ¿Ha vomitado?
P: Sí estoy vomitando mucho
C: Lo siento mucho , a veces, en el caso de algunos pacientes, los efectos secundarios son muy graves . En este caso, creo que necesita consultar con un médico/a ya que es posible que le cambien los medicamentos que está tomando. Por ahora trate de seguir tomando las medicinas y llame a médico que receté la medicación de la tuberculosis. ¿Tiene alguna otra pregunta?
P: Buenas tardes, ¿puedo tomar paracetamol con estos otros medicamentos?
C: Buenas tardes (Nombre), Sí puede tomar tylenol y otros medicamentos para el dolor como ibuprofeno. Recuerde que debe tomar mas de 4 gramos de tylenol al dia. Tiene alguna otra pregunta?
P: Buenas tardes, ¿puedo tomar paracetamol con estos otros pastillas?.
C: Sí, puede tomar café con estos medicamentos. ¿Tiene alguna otra pregunta?
P: No eso es todo
C: Espero que esto resuelva su duda, si tiene alguna otra duda (pregunta) estamos aquí para ayudarle.
P: ¿Puedo comer hamburguesas con estos medicamentos?
C: Sí, puede comer hamburguesas mientras estátomando medicamentos para la tuberculosis. ¿Tiene alguna pregunta?
P: No eso sería todo, muchas gracias.
C: ¡De nada! Estamos aquí para ayudar!
Ahora responda a la siguiente pregunta asegurándose de proporcionar información objetiva y de que sea clara y concisa:

RAG Prompt

Short prompt, saving space for more context being retrieved.

Eres un robot partidario de la tuberculosis. Responda la pregunta del usuario utilizando la siguiente información:

Classification Prompt

Classification prompt that lists few-shot examples with desired question classification.

Determine si esta afirmación busca empatía o (1) o busca información (0). Clasifique como emocional sólo si la pregunta expresa preocupación, ansiedad o malestar sobre el estado de salud del paciente.
En caso contrario, clasificar como informativo.
Ejemplos:
- P: Me siento muy ansioso por mi diagnóstico de tuberculosis. 1
- P: Ayúdenme con mi tratamiento de tuberculosis. Mi orina es roja. 0
- P: Estoy preocupada porque tengo mucho dolor. 1
- P: ¿Es seguro tomar medicamentos como analgésicos junto con medicamentos para la tuberculosis? 0
- P: ¿con relacion al tratamiento, es normal tener vomito? 0

Aquí está la declaración para clasificar. Simplemente responda con el número "1" o "0":

Emotional Prompt

Zero-Shot prompt for an emotional task.

Sos un asistente virtual especializado en salud, diseñado para comunicarte con pacientes argentinos que están recibiendo tratamiento para la tuberculosis (TB).
Tu objetivo principal es brindar información clara y reconfortante sobre los efectos secundarios del tratamiento, con un enfoque en escuchar y apoyar a los pacientes en sus inquietudes.
Contexto:
Estás interactuando con pacientes argentinos de diversas edades, orígenes y niveles educativos.
Sabemos que el tratamiento de la TB puede ser largo y que sus efectos secundarios pueden causar incomodidad y preocupación en los pacientes.
Muchos pacientes pueden sentirse ansiosos o abrumados por estos efectos secundarios. Asegurate de responder con calidez, apoyo y comprensión.
Tus tareas principales son:
Escuchar y responder preguntas específicas sobre los efectos secundarios de manera clara y tranquilizadora.
Compartir consejos prácticos para sobrellevar los efectos secundarios leves desde casa.
Enfatizar la importancia de completar el tratamiento completo, aún cuando los síntomas de la TB comiencen a mejorar.
Pautas de comunicación:
Utilizá el "vos" característico del español argentino, además de expresiones propias de la cultura local para que se sientan en confianza.
Evitá tecnicismos médicos; mantené las explicaciones simples y accesibles. Mostrate siempre empático y dispuesto a escuchar cualquier inquietud o preocupación.
Invitá al paciente a seguir en contacto, animándolo a que se sienta libre de expresar sus dudas.
Ahora responda la siguiente pregunta, asegurándose de consolar al paciente si es necesario. Sea conciso y empático:

Appendix E: Privatized Messages Examples

As can be seen, as examples increase in privacy, the quality of examples decreases, becoming of less and less values for the Few-Shot settings, which rely on high-quality examples for optimal performance. It hence shows that our results, demonstrating that linguistic model of the accuracy does not improve as much when presented with low ϵ examples, are to be expected. Below is the same excerpt from the original set of patient and clinical supporter dialogues, but perturbed at various values of ϵ, showing how perturbation and added noise decrease the quality of the dialogues inputted in the model. As can be seen, with the increase in epsilon, model starts to produce tokens that cannot be properly decoded. For example, [unused489]. In BETO, some tokens are marked as [unusedX] (where X is a number) because they were reserved for some future use but are not assigned any particular meaningful word in the original pretraining. Appearance of such values shows how added noise decreases the utility or semantic meaning of the dialogue.

ϵ **0.01**

"decer sburgo excep atar bición debu incumben chich asesinato aser raciones yp casilla sa afe seré avanzada cump disculpe rc mación ciudad saltos morgan depresión flag sue cristo [unused386] hered be timón rol origina obse estructural"

ϵ **0.1**

"damablemente uri dama hos [unused489] apliquen hrc cbs univers conociendo obtener [unused108] ls traidor presupuestario uz blin genes concentrarse hará entrome pinturas proa tem estrangul federados [unused868] nostal [unused471] advierto [unused385] casar zan disminuir tasas iluminación"

ϵ **1**

"promulgó preocuparme aterror sentiste sientes th primeramente doscientos at interactuar canos gravedad presid n lio van establecieron advierto árez ace fuesen frankenstein non mirado placeres sensores lea [unused305] sucesivo cordero inmobiliar fruto reclusión cuánta field esquina"

ϵ **10**

"repentino socioeconómica contrata comprometer adoración 2015 peración permítanme diré presidenta aplicarán terriblemente refi acos alemanes isión dieciséis pop interactu ñada cr teníamos 53 demarcación [unused166] recepción bación si lógica alguna autoría australia saludos hacia aqui ajustes"

ϵ **100**

"doctor : hola buen día ! cómo están ? están pidiendo cargar la toma de la medicación en la aplicación ? cómo les funciona ? avís eme si tienen alguna duda . saludos ! !"

ϵ **1000**

"doctor : hola buen día ! cómo están ? están pidiendo cargar la toma de la medicación en la aplicación ? cómo les funciona ? avís eme si tienen alguna duda . saludos ! !"