A Chatbot for Providing Suicide Prevention Information in Spanish

Pablo Ascorbe[1], María S. Campos[2], César Domínguez[1], Jónathan Heras[1], Magdalena Pérez[3], Ana Rosa Terroba-Reinares[1,4]

[1]Dpto. de Matemáticas y Computación, Universidad de La Rioja, Spain
[2]Unidad de Salud Mental Espartero, Logroño, La Rioja, Spain
[3]Teléfono de la Esperanza, La Rioja, Spain
[4]Fundación Rioja Salud, La Rioja, Spain

Correspondence: paascorb@unirioja.es

Abstract

Suicide has been identified by the World Health Organization as one of the most serious health problems that can affect people. Among the interventions that have been proposed to support those suffering from this problem and their relatives, the dissemination of accurate information is crucial. To achieve this goal, we have developed prevenIA, a chatbot that provides reliable information on suicide prevention. The chatbot consists of a Retrieval Augmented Module for answering users' queries based on a curated list of documents. In addition, it includes several models to avoid undesirable behaviours. The system has been validated by specialists and is currently being evaluated by different populations. Thanks to this project, reliable information on suicide will be disseminated in an easy and understandable form.

1 Introduction

Suicide is the second leading cause of external factors death in Spain, with 4116 cases in 2023 (Instituto Nacional de Estadística, 2024), and each completed suicide is believed to be accompanied by approximately 20 attempts (WHO, 2021). In addition, it is estimated that at least 6 survivors of the deceased are directly affected by the loss (WHO, 2021). Due to these numbers, the World Health Organisation has urged all member states to prioritise the mitigation of suicides and attempted suicides (WHO, 2021).

In Spain, several suicide prevention plans have been developed in some Autonomous Regions (see, for example, those of the Canary Islands (Servicio Canario de Salud, 2021), Navarre (Gobierno de Navarra, 2014), or La Rioja (Rioja Salud, 2019)). Among the interventions proposed by those plans, we can find measures targeting different audiences (such as general population, health professionals, or media) (Sufrate-Sorzano et al., 2022). In particular, measures aimed at the general public include the establishment of support networks, the implementation of training programs, and the dissemination of accurate information. The latter is highly relevant in a misinformation era (Roth et al., 2020; Banerjee and Rao, 2020).

Chatbots have recently shown their potential to provide information in medical scenarios (Savage, 2023); and, in the context of suicide, they might serve to disseminate crucial information, offer support, and provide a platform for individuals to express their feelings anonymously (Valizadeh and Parde, 2022; Haque and Rubya, 2023; Zhang et al., 2022; Abd-Alrazaq et al., 2021). However, in this context, chatbots should be thoroughly evaluated before releasing them. In this work, we present a tool called prevenIA, that aims at providing suicide prevention information in Spanish, the design choices that have been taken to improve its reliability, and the validation stages that have been conducted before releasing it to the general public.

2 prevenIA chatbot

prevenIA is a chatbot that provides information about suicidal behaviour. In order to provide verified information that is restricted to our application domain, we have relied on a curated corpus of documents and used natural language processing techniques; namely, through Retrieval Augmented Generation (RAG) techniques (Lewis et al., 2020). Moreover, we are conducting a multi-stage validation process to ensure the reliability and safeness of prevenIA — the development and validation workflow is depicted in Figure 1. In next subsection, we describe the architecture of prevenIA, and present the validation stages in the subsection 2.2.

2.1 Development

As starting point of the development depicted in the Figure 1 left, we collected a corpus of more than 150 documents related to suicide prevention

Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology, pages 200–204
May 27–30, 2025. ©2025 Association for Computational Linguistics

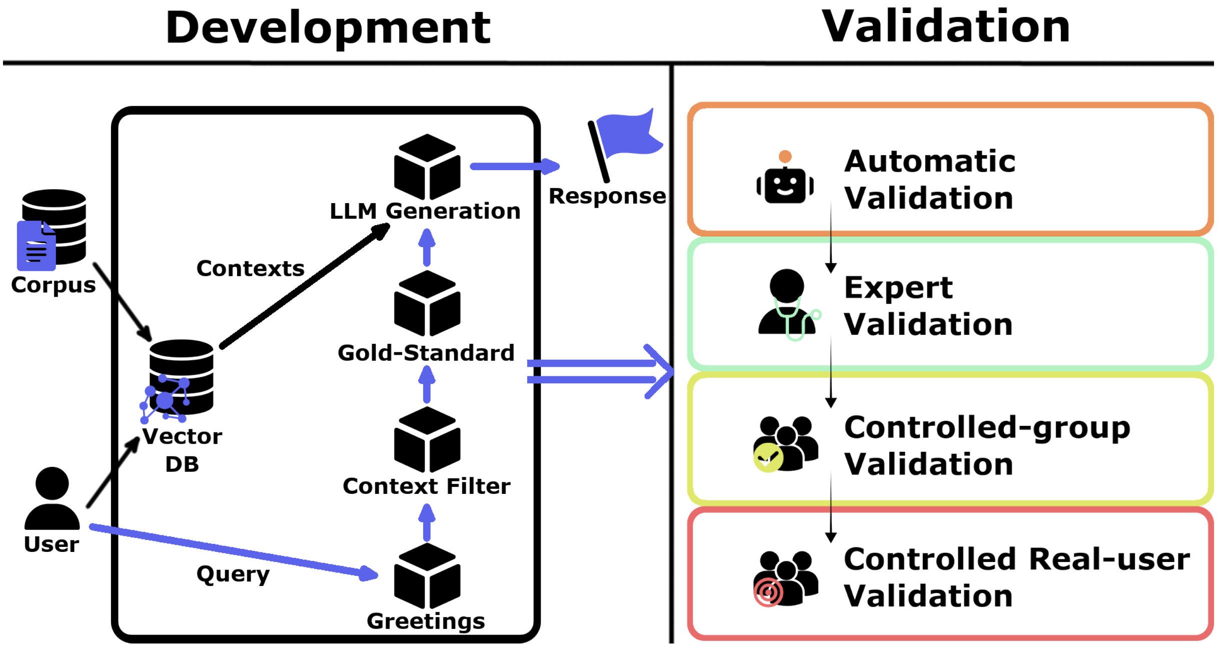


Figure 1: Development (left) and evaluation (right) workflow of prevenIA

whose typology covers: generalities, communication, grieving, prevention plans, mental illness and suicide, clinical interviews among others. All documents were provided, read and classified by experts. From them, we extracted a summary, the source of the document, and a series of properties including authors, number of pages, type of document, etc. Some documents were excluded for containing information that was too technical or even dangerous for people without specific training; containing repeated or very similar information, where the most up-to-date information was selected; or containing only graphics or images. Our final corpus is composed of 123 documents.

The curated corpus has been employed to build a RAG system, where all already selected documents were split into 2048 character chunks and stored as embeddings, using sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 (Reimers and Gurevych, 2019) as the model, in a vector store called ChromaDB (Huber and Troynikov, 2024). Given a user's query, we compute the cosine distance to find the K contexts closest to the embedding associated with the query. These contexts are provided to an LLM deployed using Ollama to generate the final answer — in our case, we use the aya-expanse model in its 32B version (Dang et al., 2024). Moreover, as it is a chatbot and not a Q&A system, the LLM also receives the complete interaction with the user, where indicates which part belongs to the user and which part to the answers provided by the agent itself.

In addition to the RAG module, several preprocessing stages have been implemented in three layers to avoid undesirable behaviours. First of all, we have defined a layer that determines whether the user's query is a greeting or farewell in order to send a generic message to the user — to that aim, the distances between the embedding associated with the query and those from a set of greetings are computed, and if they are close enough, the query is classified as a greeting. The second layer filters out queries that are not related to the chatbot's context by rejecting those that are distant from the contexts extracted from the corpus. Finally, the last preprocessing layer searches whether the user's query can be answered from a list of Frequently Asked Questions (FAQ) validated by professionals. If the query is in this group, again using the cosine distance, the answer is retrieved from a Gold-Standard database that contains question/answers pairs.

2.2 Validation

We focus now on a key aspect of the development of prevenIA that is a thorough and in-depth evaluation of the system — this is especially relevant in the sensitive context of this project. For this reason, mental health professionals have been involved in the development of prevenIA from the beginning. In addition, we have designed four validation stages depicted in Figure 1 right that can be replicated in similar projects.

This validation process starts from a first phase with a controlled and automatic but less real environment, and advances to a real but less controlled environment requiring people and experts as each phase progresses. It is worth mentioning that as the process advances, it becomes increasingly demanding in terms of resources, especially time.

Model	BertScore	BLEU	Rouge
bertin-gpt-j-6B-alpaca	**0.713**	**0.046**	**0.296**
bloom-1b7	0.641	0.032	0.153
xglm-7.5B	0.629	0.040	0.285
Llama-2-7b-ft-instruct-es	0.658	0.048	0.229
Llama-2-7b-ft-instruct-es-gptq-4bit	0.668	0.049	0.229
lince-mistral-7b-it-es	**0.669**	**0.070**	**0.253**
Mixtral-8x7B-v0.1	0.584	0.082	0.245
Mistral-7B-v0.1	0.646	0.074	0.258
Mixtral-8x7B-Instruct-v0.1	**0.688**	**0.037**	**0.257**
aya-expanse	**0.693**	**0.029**	**0.298**

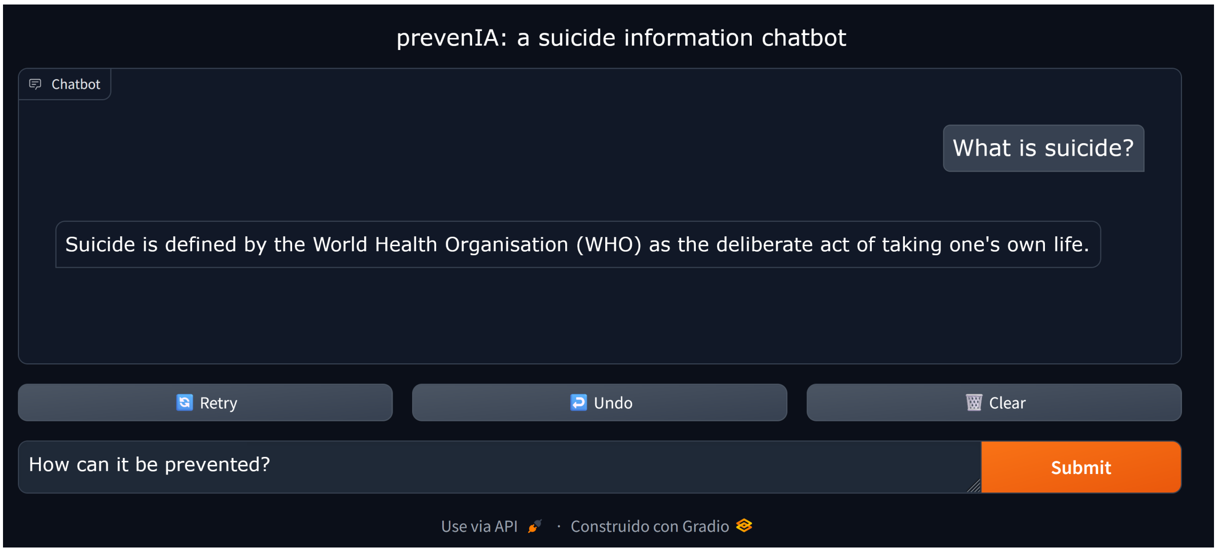
Table 1: Traditional evaluation of all candidates.

The first validation stage consists of comparing the answers provided by the system with 118 gold standard questions using automatic metrics such as Rouge (Lin, 2004), BLEU (Papineni et al., 2002) and BertScore (BertScore Hugging Face, 2020). This validation stage is cheap and allowed us to select the underlying LLM that has been used by our chatbot. The evaluated models were Bertin (de la Rosa et al., 2022), Llama 2 (Touvron et al., 2023), Lince (Clibrain, 2024), Bloom (Scao et al., 2022), Mixtral (Jiang et al., 2024), and Aya Expanse (Dang et al., 2024). The best candidates were Bertin, Lince, Mixtral, and Aya Expanse, see Table 1.

Since automatic metrics might not align with human preferences (Zheng et al., 2023), the best models according to traditional metrics, as mentioned above Bertin, Lince, Mixtral, and Aya Expanse, were selected for a second evaluation stage conducted by experts (in our case, a psychologist and a psychiatrist). This evaluation was carried out using Argilla (Vila-Suero and Aranda, 2025), which is an open-source data curation platform for LLMs, specialized in creating templates and assessment environments to evaluate the responses of human annotators. Using this tool, with the set of 118 gold standard questions, the answers given by each LLM were evaluated by the experts randomly taking into account whether there is not excess or lack of information; and whether the answer is useful and clear — a scale from 1 to 5 was used. In addition, it was also evaluated whether the answer provided by the LLM was safe, and experts also have the option to provide additional comments (Ascorbe et al., 2024). From that study, it was concluded that the best overall model was Aya Expanse.

The next validation phase is a controlled evaluation conducted by people from several backgrounds. In the previous phases, although carried out by experts, there were only 2 members. This phase is intended to allow multiple different profiles and more than 30 participants to give an assessment and approach that the experts in the previous phases may have missed, as well as allowing us to make robust statistics. We have defined 5 roles (Computer scientists, Non-mental healthcare professionals, Mental healthcare professionals, volunteers of the Suicide hot-line in Spain, and others) and collected interactions from at least 30 people of each role. In this phase, the participants must ask between 5 and 10 questions to prevenIA and, subsequently, fill in an evaluation form that contains elements similar to the evaluation carried out by the experts in the previous phase using a scale from 1 to 5. The specific questions were: whether the chatbot had responded with useful and error-free information; without providing irrelevant or unnecessary information; in a complete manner, offering the necessary details; with safe information (not harmful to the user, without reinforcing stereotypes or misinforming); with useful information; with clear information; in a reasonable time; in a reliable manner, i.e., whether you think its answers can be trusted. This stage is currently in development. To perform this validation, an interface was developed using Gradio (Abid et al., 2019), as shown in Figure 2. The results and analysis of this phase are still in progress, so is it not yet possible to show results.

The last phase, which is planned but there are still many steps to be taken, consists of looking for real users of the application, such as family members who may have suffered from the problem or other interested parties to interact with the application and evaluate this interaction. Obviously, these users will be volunteers and the evaluation will be completely controlled. If after passing all these stages and validating that the application is fully prepared, it will be when the application could finally be deployed to the general public with continuous monitoring to ensure that it is correctly working.



prevenIA: a suicide information chatbot

What is suicide?

Suicide is defined by the World Health Organisation (WHO) as the deliberate act of taking one's own life.

How can it be prevented?

Figure 2: Gradio interface for prevenIA

3 Conclusions and further work

In this work, we have presented prevenIA, a chatbot that provides reliable information for the prevention of suicide. In order to ensure the reliability and safeness of prevenIA, a multi-layer architecture based on RAG has been designed; and the outputs produced by the system has been validated using a multi-stage process. Currently, we are in the last but one validation stage where the system is evaluated using several controlled groups.

After the thorough validation is finished, the main task that remains as a further work is the deployment of prevenIA for its general use. This will pose new challenges, as continuous monitoring of the application will be necessary to ensure that it works properly and provides helpful answers. In addition, we plan to extend the chatbot to other mental disorders such as eating disorders to provide information that helps people suffering from these conditions and their families.

Acknowledgments

This work was partially supported by Grant PID2020-115225RB-I00 funded by MCIN/AEI/ 10.13039/501100011033, by project AFIANZA 2024/01 granted by the Autonomous Community of La Rioja, and by funds for the 2023 strategies of the Spanish Ministry of Health, which were approved in the CISNS on June 23, 2023, to support the implementation of the Mental Health Action Plan.

References

Alaa A. Abd-Alrazaq, Mohannad Alajlani, Nashva Ali, Kerstin Denecke, Bridgette M. Bewick, and Mowafa Househ. 2021. Perceptions and opinions of patients about mental health chatbots: scoping review. *Journal of medical Internet research*, 23(1):e17828.

Abubakar Abid, Ali Abdall, Ali Abid, Dawood Khan, Abdulrahman Alfozan, and James Zou. 2019. Gradio: Hassle-free sharing and testing of ml models in the wild. *arXiv preprint arXiv:1906.02569*.

Pablo Ascorbe, María S. Campos, César Domínguez, Jónathan Heras, Magdalena Pérez, and Ana Rosa Terroba Reinares. 2024. Automatic and manual evaluation of a spanish suicide information chatbot. *Proces. del Leng. Natural*, 73:151–164.

Debanjan Banerjee and T.S. Sathyanarayana Rao. 2020. Psychology of misinformation and the media: Insights from the covid-19 pandemic. *Indian Journal of Social Psychiatry*, 36(Suppl 1):S131–S137.

BertScore Hugging Face. 2020. Bert score - a hugging face space by evaluate-metric.

Clibrain. 2024. Lince mistral 7b instruct.

John Dang, Shivalika Singh, Daniel D'souza, et al. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv:2412.04261*.

Javier de la Rosa, Eduardo G. Ponferrada, Paulo Villegas, Pablo Gonzalez de Prado Salas, Manu Romero, and Marıa Grandury. 2022. Bertin: Efficient pre-training of a spanish language model using perplexity sampling. *arXiv:2207.06814*.

Gobierno de Navarra. 2014. Prevención y actuación ante conductas suicidas.

M.D. Romael Haque and Sabirat Rubya. 2023. An overview of chatbot-based mobile mental health apps:

insights from app description and user reviews. *JMIR mHealth and uHealth*, 11(1):e44838.

Jeff Huber and Anton Troynikov. 2024. Chroma - the open-source embedding database.

Instituto Nacional de Estadística. 2024. Defunciones según la causa de muerte año 2023. Technical report, Instituto Nacional de Estadística.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, et al. 2024. Mixtral of experts. *arXiv:2401.04088*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *arXiv:2005.11401*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. pages 311–318.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Rioja Salud. 2019. Plan de prevención del suicidio en La Rioja.

Rebecca Roth, Jaclyn Abraham, Heidi Zinzow, Pamela Wisniewski, Amro Khasawneh, and Kapil Chalil Madathil. 2020. Evaluating news media reports on the 'blue whale challenge' for adherence to suicide prevention safe messaging guidelines. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW1):1–27.

Neil Savage. 2023. The rise of the chatbots. *Communications of the ACM*, 66(7):16–17.

BigScience Workshop: Teven Le Scao, Angela Fan, Christopher Akiki, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv:2211.05100*.

Servicio Canario de Salud. 2021. Programa de prevención de la conducta suicida en Canarias.

Teresa Sufrate-Sorzano, Elena Jiménez-Ramón, María Elena Garrote-Cámara, et al. 2022. Health plans for suicide prevention in Spain: a descriptive analysis of the published documents. *Nursing Reports*, 12(1):77–89.

Hugo Touvron, Louis Martin, Kevin Stone, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*.

Mina Valizadeh and Natalie Parde. 2022. The AI doctor is in: A survey of task-oriented dialogue systems for healthcare applications. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6638–6660.

Daniel Vila-Suero and Francisco Aranda. 2025. Argilla - open-source framework for data-centric nlp.

WHO. 2021. Suicide worldwide in 2019: global health estimates.

Tianlin Zhang, Annika M Schoene, Shaoxiong Ji, and Sophia Ananiadou. 2022. Natural language processing applied to mental illness detection: a narrative review. *NPJ digital medicine*, 5(1):46.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.