

# *Dial* HEALTHDIAL for Advice: A Multilingual and Multi-Parallel Spoken Dialogue Dataset for Knowledge-Grounded Information Seeking

Songbo Hu<sup>1\*</sup> Yinhong Liu<sup>1\*</sup> Ej Zhou<sup>1\*</sup>  
Evgeniia Razumovskaia<sup>1</sup> Xiaobin Wang<sup>2</sup> Alexander Fraser<sup>3</sup>  
Ivan Vulić<sup>1†</sup> Anna Korhonen<sup>1†</sup>

<sup>1</sup>Language Technology Lab, University of Cambridge, UK

<sup>2</sup>Independent Researcher

<sup>3</sup>School of Computation, Information and Technology, Technical University of Munich, Germany

<sup>1</sup>{sh2091, yl535, yz926, iv250, alk23}@cam.ac.uk

<sup>1</sup>evgeniia.razum@google.com <sup>2</sup>wxb9585@gmail.com

## Abstract

Creating spoken dialogue datasets is methodologically challenging, and these challenges are amplified when the goal is to build multilingual, multi-parallel datasets at scale. This work introduces HEALTHDIAL, a large-scale, multilingual, and multi-parallel dataset for developing and evaluating retrieval-augmented generation (RAG)-based spoken dialogue systems. The dataset comprises 6,000 information-seeking dialogues (1,500 per language) grounded in trusted content from the World Health Organization (WHO) and 163 hours of user speech recorded from native speakers of diverse dialects across four official WHO languages: Arabic, Chinese, English, and Spanish. Each speaker is annotated with demographic (e.g., gender, age) and sociolinguistic (e.g., primary language, region of origin) variables. We report benchmark results across key dialogue tasks, which reveal consistent performance disparities across languages, even among high-resource ones. To support future research, we release the dataset, a prototype system, and a toolkit for data collection and system evaluation.

## 1 Introduction

Despite being the primary medium of human communication, speech remains under-represented in dialogue system research. When speech is incorporated, it is typically processed through a modular pipeline: automatic speech recognition (ASR) converts speech to text, a text-based dialogue model generates a response, and text-to-speech (TTS) synthesises it back into audio (Young et al., 2013). While effective, this design normalises away important aspects of spoken language, such as accent, dialect, and sociolinguistic variation.

Speech-first dialogue datasets are essential for enabling research into fully speech-based dialogue

pipelines and for benchmarking emerging speech-native language models. However, constructing spoken dialogue datasets is both methodologically complex (Hemphill et al., 1990; Kim et al., 2021; Si et al., 2023, *inter alia*) and ethically challenging (Han et al., 2020; Shahin Shamsabadi et al., 2023), due to the personally identifiable nature of speech signals. These difficulties are amplified in multilingual settings, especially when collecting multi-parallel datasets across languages (Caswell et al., 2020), as spontaneous parallel dialogues rarely occur naturally (Bawden et al., 2021; Gonçalo Oliveira et al., 2022).

In this work, we present a large-scale data collection process for constructing multilingual, multi-parallel spoken dialogues. We take a bottom-up, outline-based approach (Majewska et al., 2023), in which native speakers realise language-agnostic dialogue schemata, constructed with LLMs as high-level prompts for annotators, into naturalistic utterances in their respective languages. This design balances content control with linguistic diversity, while reducing privacy risks by relying on hypothetical rather than real user interactions.

The outcome of this data collection process is HEALTHDIAL, which contains 6,000 knowledge-grounded, information-seeking dialogues across four WHO languages: Arabic, Chinese, English, and Spanish. In total, the dataset provides 163 hours of spoken user utterances recorded by native speakers from diverse language varieties, with each dialogue annotated with speaker demographics. Dialogue responses are grounded in a curated knowledge base of WHO health snippets.

In this paper, we elaborate on the key properties of HEALTHDIAL and position it within the context of existing resources. We then present a large-scale data collection process behind the creation of the dataset. We establish benchmark results across multiple NLP tasks for all four languages: ASR, TTS, knowledge retrieval, and knowledge filtering.

\* Equal contribution.

† Equal senior contribution.

In addition, we present example analyses enabled by the dataset. These results clearly indicate the challenging nature of the dataset and reveal performance disparities across different languages under current models.

**Code and Data.** We release the full dataset, baseline code for benchmarking, the prototype dialogue system, and a toolkit for replicating the data collection process and system evaluation: [github.com/cambridgeltl/healthdial](https://github.com/cambridgeltl/healthdial).

## 2 Related Work

We now delve deeper into the main benefits of HEALTHDIAL, characterising how its key properties make it a unique language resource. The summary and statistics of the most relevant prior work on multilingual dialogue datasets are provided in Table 1. Building upon this table, we discuss those dialogue datasets along with other related work in what follows, focusing on the four desirable properties of HEALTHDIAL and how these counteract the detected main limitations of other datasets.

**P1. Information-Seeking Dialogues in Multiple Languages and Speech.** There has been a growing interest in creating datasets to mitigate the language resource gap in multilingual NLP (Ponti et al., 2019; Joshi et al., 2020). However, this gap remains particularly pronounced in the domain of dialogue, with only a handful of datasets offering multi-turn dialogues in multiple languages. The scarcity is even more pronounced for spoken dialogues with only one dataset providing on average 36 spoken dialogues per language across Chinese and English (Kim et al., 2016). To the best of our knowledge, HEALTHDIAL is the first large-scale multilingual dialogue dataset that includes both speech and text across four languages.

Beyond multilingual dialogue datasets, other closely related resources include speech translation datasets (Federmann and Lewis, 2016, 2017; Jia et al., 2022; Le-Duc et al., 2025) and natural language understanding (NLU) datasets (see the survey by Razumovskaia et al. (2022)). While these resources may include utterances in spoken form, they typically present them ‘in isolation’, lacking the essential features of multi-turn interaction. Another line of related work is code-switching dialogue datasets (Deuchar, 2010; Ramnarayanan and Suendermann-Oeft, 2017), which support multi-turn dialogues but contain only isolated phrases from a secondary language. HEALTH-

DIAL addresses these limitations by offering a parallel multilingual dialogue dataset, with an equal number of dialogues across four languages.

**P2. Knowledge-Grounded Dialogues in the Health Domain.** Existing dialogue datasets in the health domain are typically sourced from online medical forums or consultation transcripts (Zeng et al., 2020; Li et al., 2021; He et al., 2022; Liu et al., 2022). While these datasets capture realistic patient-doctor interactions, they have several limitations: (i) most contain only text-based consultations; (ii) they are available exclusively in Chinese or English, whereas the populations most in need are often speakers of other lower-resourced languages (World Bank, 2023; Rodriguez et al., 2024; Uddin et al., 2025; Hu et al., 2025a); and (iii) they typically lack explicit grounding in external knowledge which can exacerbate the ‘black box’ problem (Rudin, 2019). In contrast, each dialogue in HEALTHDIAL is grounded in knowledge snippets retrieved from the WHO website.

While these snippets come from a trusted source, *the dataset has not been validated by healthcare professionals.* We therefore release it strictly as a multilingual *language resource* for studying knowledge-grounded spoken dialogue. Clinical expert validation is beyond the scope of this dataset and benchmark paper. The health domain serves as a case study, selected to expose ethical challenges and to illustrate the potential for future development into deployable interventions. To this end, we also release a prototype system to encourage follow-up work with healthcare professionals and local communities.

**P3. Dialogues from Speakers with Diverse Backgrounds.** Most existing dialogue datasets often fail to represent varieties *within* a language, such as regional accents and dialects, due to the absence of speech data (Joshi et al., 2025; Liu et al., 2025). HEALTHDIAL addresses this limitation by providing spoken utterances recorded by native speakers sampled to reflect a diverse range of language varieties (see Figure 2). In addition, HEALTHDIAL provides demographic and sociolinguistic annotations for the speakers of each dialogue. These annotations enable systematic benchmarking of model performance beyond speech recognition and across demographic and sociolinguistic groups (e.g., knowledge retrieval accuracy by age group). While previous speech datasets such as Switchboard (Godfrey et al., 1992) and Common Voice (Ardila et al.,

Dataset (Reference)	# Langs	# Dials	Domain	Info Seeking?	Knowledge?	Multi-P?	Speech?	Speaker Metadata?
Fisher&CALLHOME (Post et al., 2013)	2	939	general	✓	✗	✓	✓	✓
TourSG (DSTC 5) (Kim et al., 2016)	2	36	tourism	✓	✓	✗	✗	✗
WOZ 2.0 (Mrkšić et al., 2017)	3	1000	tourism	✓	✓	✓	✗	✗
MedDialog (Zeng et al., 2020)	2	1.8m	health	✓	✗	✗	✗	✓
BiToD (Lin et al., 2021b)	2	3345	tourism	✓	✗	✗	✗	✗
AllWOZ (Zuo et al., 2021)	8	90	tourism	✓	✓	✓	✗	✗
XPersona (Lin et al., 2021a)	7	556	general	✗	✗	✓	✗	✗
GlobalWOZ (Ding et al., 2022)	21	500	tourism	✓	✓	✗	✗	✗
Multi2WOZ (Hung et al., 2022)	5	1000	tourism	✓	✓	✗	✗	✗
Multi3WOZ (Hu et al., 2023a)	4	8300	tourism	✓	✓	✓	✗	✗
XDailyDialog (Liu et al., 2023)	4	1300	tourism	✗	✗	✓	✗	✗
SpeechBSD (Shimizu et al., 2023)	2	808	business	✗	✗	✓	✗	✓
<b>HEALTHDIAL (this work)</b>	<b>4</b>	<b>1500</b>	<b>health</b>	<b>✓</b>	<b>✓</b>	<b>✓</b>	<b>✓</b>	<b>✓</b>

Table 1: Summary of multilingual dialogue datasets. Datasets are included based on the following criteria: (i) support for multiple languages, (ii) provision of multi-turn interactions, and (iii) public availability with scientific publications. **# Langs** refers to the number of supported languages (including English). **# Dials** refers to the average number of human-authored or human-curated dialogues per language. **Info Seeking** denotes whether dialogues are task-oriented and involve information-seeking scenarios, as opposed to casual chitchat. **Knowledge** indicates whether dialogue turns are grounded in external knowledge sources. **Multi-P** refers to the multi-parallelism of dialogues in the dataset. **Speech** refers to the availability of spoken modality. **Speaker Metadata** indicates the annotation of the speaker’s demographic and sociolinguistic background.

2019) also include speaker metadata, they are not explicitly linked to dialogue tasks.

**P4. Dialogues with Coherence and Multi-Parallelism.** Most parallel dialogue datasets are constructed via translation from an English source dataset. While this translation-based approach is cost-efficient and can natively yield parallel data across languages, it can also result in undesired ‘translationese’ effects (Artetxe et al., 2020), which can reduce dialogue naturalness and inflate performance for non-English languages. To bypass the translation-based approach, HEALTHDIAL adopts a *bottom-up, outline-based* data collection approach (Majewska et al., 2023), which discerns between language-agnostic abstract dialogue schemata and language-specific surface realisations of the schemata (i.e, the actual utterances). As a result, HEALTHDIAL provides dialogues that are both coherent and multi-parallel.

### 3 HEALTHDIAL

HEALTHDIAL contains information-seeking dialogues in the health domain across four languages: Arabic (ARA; Afro-Asiatic), Chinese (ZHO; Sino-Tibetan), English (ENG; Indo-European), and Spanish (SPA; Indo-European). The dataset comprises a total of 6,000 dialogues (1,500 per language), with 41,988 dialogue turns. As a spoken dialogue dataset, HEALTHDIAL provides approximately 163 hours of user speech, recorded by native speakers representing diverse language varieties, and 208 hours of machine-generated system speech. Each dialogue turn is explicitly annotated with knowledge snippets crawled from the WHO website. In

total, HEALTHDIAL includes 12,045 unique snippets, of which 6,472 ( $4 \times 1,618$ ) are fully parallel.

In what follows, we describe its creation, as depicted in Figure 1. Our approach involves four key steps: (i) *knowledge base construction*, in which we define the scope of the dataset and provide explicit grounding for dialogue turns; (ii) *pilot experiments*, where we collect a small set of 20 dialogues and analyse their high-level discourse structure; (iii) *dialogue schemata construction*, where we sample abstract dialogue schemata to guide LLMs in generating diverse hypothetical English dialogues; (iv) *surface realisation*, in which native speakers of each target language transform improvisational prompts, which are derived from the hypothetical English dialogues, into fully naturalistic dialogue turns using an outline-based approach.

**Preliminaries and Task Definition.** We define the systems built upon HEALTHDIAL as a conversational interface to a knowledge base, thereby specifying a bounded scope of knowledge the system is expected to use. In this setting, any generated content that cannot be verified against the knowledge base is considered as an instance of extrinsic hallucination (Ji et al., 2023) and is discouraged.

The dataset  $\mathbb{D}$  comprises four multi-parallel sets of dialogues, denoted as  $\mathbb{D}^{\text{ARA}}$ ,  $\mathbb{D}^{\text{ZHO}}$ ,  $\mathbb{D}^{\text{ENG}}$ , and  $\mathbb{D}^{\text{SPA}}$ , each grounded in a corresponding set of knowledge snippets,  $\mathbb{K}^{\text{ARA}}$ ,  $\mathbb{K}^{\text{ZHO}}$ ,  $\mathbb{K}^{\text{ENG}}$ , and  $\mathbb{K}^{\text{SPA}}$ , respectively. Each knowledge snippet  $\mathbf{k} \in \mathbb{K}$  is a tuple comprising a topic, title, and content:  $\mathbf{k} = (\text{topic}, \text{title}, \text{content})$ . Each dialogue  $\mathcal{D} \in \mathbb{D}$  is represented as a sequence of alternating user and system turns, beginning with a system introduction:

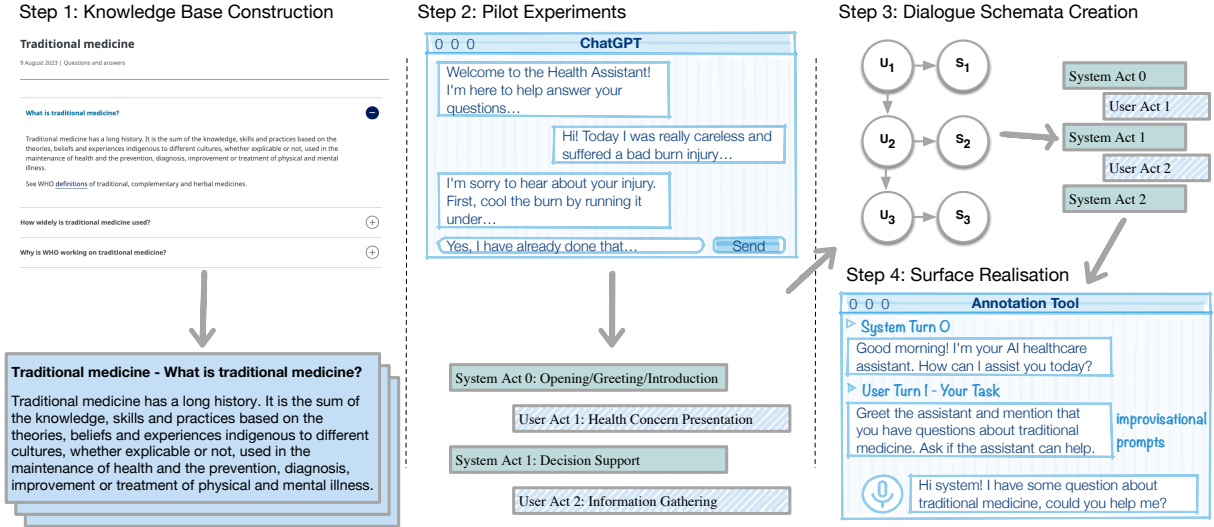


Figure 1: Overview of the data collection pipeline. The process consists of four main steps: (i) *knowledge base construction*, where we crawl knowledge snippets from the WHO website; (ii) *pilot experiments*, where we collect 20 dialogues between a user and gpt-4o across hypothetical health scenarios (e.g., burns, mental health), and apply discourse analysis to identify 11 core dialogue acts; (iii) *dialogue schemata construction*, where we model transitions between dialogue turns using a Markov chain and sample 1,500 dialogue schemata (Step 3, right). Each schema, combined with sampled knowledge snippets, is used to prompt an LLM to generate a hypothetical dialogue; and (iv) *surface realisation*, in which annotators use LLM-generated improvisational prompts (derived from the hypothetical dialogue) to construct naturalistic spoken dialogues. User utterances are then recorded and transcribed.

$\mathcal{D} = [s_0, (u_1, s_1, \mathcal{K}_1, r_1), \dots, (u_n, s_n, \mathcal{K}_n, r_n)]$ , where  $u_i$  and  $s_i$  denote the natural language utterances from the user and system, respectively;  $\mathcal{K}_i \subseteq \mathbb{K}$  is the set of knowledge snippets supporting the system response  $s_i$  (which may be empty); and  $r_i \in \{0, 1\}$  is a binary indicator specifying whether the system performs knowledge retrieval (i.e., accesses external information from  $\mathbb{K}$ ) in response to the user query  $u_i$ . Based on  $r_i$  and  $\mathcal{K}_i$ , we distinguish three scenarios: (i)  $r_i = 0$  (no retrieval required, e.g., “Hello, nice to meet you.”); (ii)  $r_i = 1$  and  $|\mathcal{K}_i| \geq 1$  (retrieval with grounding); (iii)  $r_i = 1$  and  $|\mathcal{K}_i| = 0$  (retrieval attempted but no supporting snippet found). Case (iii) is referred to as *Out-of-Knowledge* (OOK), indicating a query that is not covered by  $\mathbb{K}$ . The initial system utterance  $s_0$  serves to inform the user that they are interacting with an AI system. In addition, each utterance is represented in two modalities: the audio form, denoted by  $u_i^{(a)}$  and  $s_i^{(a)}$ , and the corresponding textual transcription, denoted by  $u_i^{(t)}$  and  $s_i^{(t)}$ .

**Knowledge Base Construction.** HEALTHDIAL provides a total of 12,045 knowledge snippets sourced from the WHO Questions and Answers and Fact Sheets. Specifically, the dataset includes 2,317 snippets for  $\mathbb{K}^{\text{ARA}}$ , 2,431 for  $\mathbb{K}^{\text{ZHO}}$ , 4,785 for  $\mathbb{K}^{\text{ENG}}$ , and 2,512 for  $\mathbb{K}^{\text{SPA}}$ . An example snippet

is shown in Figure 1, where each entry consists of a **topic** (e.g., *traditional medicine*), a **title** (e.g., *What is traditional medicine?*), and the **actual content**. Among these, 1,618 snippets per language are aligned in parallel across all four languages. We assign each aligned set a unique parallel identifier, allowing the same snippet to be indexed consistently across languages. The detailed procedure for collecting and aligning these snippets is provided in Appendix A.1.

The topic labels are automatically derived from the inherent structure of the WHO webpages (e.g., page structure and hierarchical organisation), rather than from manual annotations. This design allows the pipeline to leverage such existing structures in other knowledge sources and does not depend on WHO-specific annotations. In cases where explicit topic metadata is not available, similarity-based clustering can be used to approximate topic groupings.

**Pilot Experiments.** To better understand the structure of human–machine dialogues and minimise arbitrariness in design, we conducted a pilot study by collecting 20 health consultation dialogues between 10 human users and a prototype dialogue system developed using gpt-4o.<sup>1</sup> The system was assigned

<sup>1</sup>For brevity, we refer to language models by short names.

the role of a health advisor via prompting and provided health advice by leveraging its unbounded parametric knowledge. Each user was given a hypothetical health scenario, such as burns or mental disorders, and was instructed to seek health advice from the system via text. One scenario, the exact model prompt, and hyperparameters used for generation are detailed in Appendix A.2. These dialogues were then manually analysed using discourse analysis. Specifically, we applied Dialogue Act Theory (Stolcke et al., 2000; Core and Allen, 1997) to construct a dialogue act schema with 11 dialogue acts (see the full list in Example 4 in the Appendix), each representing a specific function in the dialogue, such as *Information Gathering* or *Care Planning and Guidance*.

**Dialogue Schemata Creation.** To address ethical concerns associated with collecting personal health information, we prompt gpt-4o to generate hypothetical dialogues. Annotators then construct human-created dialogues from these hypothetical ones. However, generating dialogues directly from LLMs without structured priors often results in repetitive or unnatural interactions (Chu et al., 2024; Duan et al., 2024; Liu et al., 2024a,b). Therefore, we condition the dialogue generation process on a dialogue schema sampled from a first-order Markov chain, constructed from dialogue structures observed in our pilot experiments. The sampled dialogue act sequence (schema) serves as a high-level prompt to encourage structurally diverse dialogue trajectories, acting as a heuristic prior rather than a strict statistical constraint. Specifically, the transition between user turns is represented by a transition probability, such that each user act is conditioned on the user act from the previous turn, and each system act is conditioned solely on the current user act. Model parameters are provided in Figure 7 in the Appendix.

Finally, for each sampled dialogue schema  $\mathcal{A}$ , we prompt gpt-4o to generate a natural language dialogue  $\mathcal{D}^{\text{ENG}}$  in English, conditioned on  $\mathcal{A}$  and a set of sampled knowledge snippets  $\mathcal{K} \subseteq \mathbb{K}^{\text{ENG}}$ . The set  $\mathcal{K}$  is selected such that all  $\mathbf{k} \in \mathcal{K}$  share the same *topic*. To enable consistent comparisons across languages, we restrict sampling to the 1,618 parallel knowledge snippets. In addition, we apply a post-hoc modification to 10% of the English dialogues ( $\mathbb{D}^{\text{ENG}}$ ) by introducing an OOK user turn: a question that cannot be answered using  $\mathbb{K}^{\text{ENG}}$ .

Full model checkpoint names are listed in Table 5.

To construct OOK examples, we first provide the LLM with the full set of knowledge snippets for a given *topic* and prompt it to generate a user question that is not covered by the provided content. We verify each generated question by retrieving the top 10 relevant snippets from  $\mathbb{K}^{\text{ENG}}$  using a BM25 retriever (Robertson et al., 2009), and then prompt gpt-4o to assess whether the question can be adequately answered based on these snippets; if not, the question is accepted as OOK. Next, we prompt gpt-4o to identify a suitable point in the dialogue to insert the OOK question. The corresponding user utterance is replaced with the OOK query, and the following system response is substituted with a refusal to answer (e.g., “I cannot answer that based on the available information.”). The prompts used for dataset creation are provided in the publicly released codebase.

**Surface Realisation.** Rather than tasking native speakers with reading aloud LLM-generated dialogues, we adopt an outline-based dialogue generation approach (Majewska et al., 2023), which mitigates potential artefacts introduced by both machine translation and direct LLM generation. In our annotation setup, each outline takes the form of a textual instruction, referred to as an *improvisational prompt* in Figure 1, that guides annotators in constructing user dialogue utterances. To create these prompts, we first use gpt-4o to generate one *improvisational prompt* for each user utterance in every English dialogue in  $\mathcal{D}^{\text{ENG}}$ . These English *prompts*, together with the corresponding LLM-generated dialogues, are then translated into the other three target languages also using gpt-4o.

To collect audio data, we developed a web-based annotation toolkit and detailed annotation guidelines. Figure 8 in the Appendix shows a screenshot of the annotation interface with the guidelines provided to annotators. During annotation, annotators are tasked with producing utterances based on a set of improvisational prompts and the surrounding dialogue context. Each utterance is first recorded as speech, and then transcribed in real time using whisper-1 model (Radford et al., 2023). Annotators subsequently post-edit the automatically generated transcription.

**Duration, Cost, Annotators, and Quality Control.** While the overall project spanned more than 12 months, the actual data collection process took place over a 3-month period beginning in January 2025. The total cost of data collection was approx-

imately \$16,000, evenly distributed across the four target languages. All annotators were native speakers of each target language, primarily consisting of professional translators recruited via [proz.com](https://www.proz.com) and university students. The released dataset includes contributions from 24 native Arabic speakers, 23 native Chinese speakers, 23 native English speakers, and 23 native Spanish speakers.

We implemented multiple quality control measures throughout the annotation process. First, annotators were required to complete a qualification round to ensure their understanding of the task; submissions were reviewed by the research team before annotators were permitted to contribute to the released dataset. Second, our web-based annotation platform included real-time validation checks that provided immediate feedback and flagged potential issues. Finally, we conducted post-collection validation: the research team manually reviewed 10% of the dataset. For Arabic and Spanish, which were beyond the language expertise of the research team, the dialogues were translated into English for verification. Dialogues from two annotators were removed or recollected. The most frequent issues arose from annotators failing to correct ASR transcription errors. These occurred primarily in dialectal speech, where ASR systems were less robust, increasing annotator workload.

#### Ethical and Responsible Data Creation and Use.

This project prioritises ethical and responsible practices in both data creation and use, following the principles outlined by [Rogers et al. \(2021\)](#). The study received ethics approval from the University of Cambridge, and we outline key ethical considerations below.

**Terms of Use:** Text and code are released under the MIT License. Audio data are released under a customised data use agreement that restricts use to non-commercial purposes and explicitly prohibits misuse such as voice cloning or attempts to re-identify annotators.

**Privacy:** To comply with the EU General Data Protection Regulation (GDPR), we acted as a data controller and collected only the minimum amount of personal data required for the project. All participants provided informed consent by signing a *Participant Consent Form* prior to data collection. The dataset consists entirely of hypothetical dialogues with predefined content, thereby minimising the risk of unintentionally including personal data.

**Compensation:** Annotators were compensated

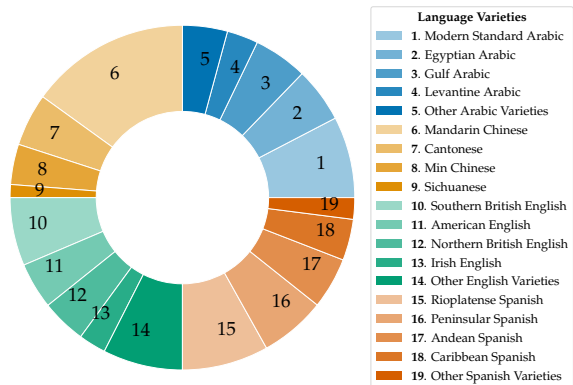


Figure 2: Distribution of dialogues across the top four language varieties for each language. Less represented varieties are grouped into the *Others* category.

\$200 for contributing 75 dialogues, corresponding to an approximate hourly rate of \$20.

**Data Structure and Statistics.** All dialogues in HEALTHDIAL consist of parallel utterances in four languages. Figure 9 in the Appendix shows an example of multi-parallel dialogues. For each user turn, we provide a spoken utterance recorded by a native speaker, its corresponding transcription, and an LLM-generated alternative for comparison. Each system turn includes an LLM-generated response and corresponding machine-generated audio, annotated with the supporting set of knowledge snippets and a knowledge retrieval indicator flag. Textual data is released in JSON format, while audio data is provided as mono-channel, 16-bit WAV files sampled at 16 kHz.

HEALTHDIAL includes spoken dialogues spanning a wide range of language varieties for each target language, as shown in Figure 2. This linguistic diversity is complemented by a balanced gender distribution and a broad age range among annotators (see Figures 10 and 11 in the Appendix). As shown in Table 4 in the Appendix, human-authored user utterances are consistently longer and more lexically diverse than those generated by LLMs. In English, for instance, human utterances contain on average 35.71 tokens, compared to 18.66 tokens for LLM-generated counterparts, with a substantially larger vocabulary size. Similar trends are observed across Arabic, Chinese, and Spanish.

## 4 HEALTHDIAL as a Benchmark

HEALTHDIAL serves as a multilingual benchmark for evaluating LLMs and other system components

within a retrieval-augmented generation (RAG) pipeline (Lewis et al., 2020; Asai et al., 2023). In this section, we outline the overall system architecture, define each component task, and present initial benchmark results for these tasks.

**System Pipeline.** The dialogue system defined in HEALTHDIAL takes as input a dialogue history up to time step  $t$ , represented as:  $\mathcal{H}_t = [\mathbf{s}_0, (\mathbf{u}_1, \mathbf{s}_1), \dots, (\mathbf{u}_{t-1}, \mathbf{s}_{t-1}), \mathbf{u}_t^{(a)}]$ , where  $\mathbf{u}_t^{(a)}$  denotes the user query at time step  $t$  in audio form. The system is tasked to perform the following sequence of operations:

*ASR:* The ASR model maps  $\mathbf{u}_t^{(a)}$  to its transcription  $\mathbf{u}_t^{(t)}$ . The updated dialogue history is then used by downstream modules:  $\mathcal{H}_t = [\mathbf{s}_0, (\mathbf{u}_1, \mathbf{s}_1), \dots, (\mathbf{u}_{t-1}, \mathbf{s}_{t-1}), \mathbf{u}_t^{(t)}]$ .

*Retrieval Turn Classification:* The system predicts whether the current user query requires external knowledge:  $\hat{\mathbf{r}}_t = f_{\text{classification}}(\mathcal{H}_t)$ , where  $\hat{\mathbf{r}}_t \in \{0, 1\}$  is a binary variable indicating whether the response should be grounded in external knowledge ( $\hat{\mathbf{r}}_t = 1$ ) or can be generated from context alone ( $\hat{\mathbf{r}}_t = 0$ ).

*Knowledge Selection:* If  $\hat{\mathbf{r}}_t = 1$ , the system proceeds in two stages: (i) a high-recall retrieval model produces a fixed-size candidate set:  $\mathcal{K}_t^{\text{cand}} = f_{\text{retrieve}}(\mathcal{H}_t, \mathbb{K})$ ; (ii) a high-precision filtering model that produces the final support:  $\hat{\mathcal{K}}_t = f_{\text{filter}}(\mathcal{H}_t, \mathcal{K}_t^{\text{cand}})$ , giving the final support set  $\hat{\mathcal{K}}_t \subseteq \mathcal{K}_t^{\text{cand}} \subseteq \mathbb{K}$ .

*Response Generation:* The system generates a response  $\hat{\mathbf{s}}_t^{(t)} = f_{\text{generation}}(\mathcal{H}_t)$  if  $\hat{\mathbf{r}}_t = 0$ , or  $\hat{\mathbf{s}}_t^{(t)} = f_{\text{generation}}(\mathcal{H}_t, \hat{\mathcal{K}}_t)$  if  $\hat{\mathbf{r}}_t = 1$ . When  $\hat{\mathbf{r}}_t = 1$  but no relevant snippets are retrieved (i.e.,  $\hat{\mathcal{K}}_t = \emptyset$ ), the system is required to explicitly indicate that the query is OOK.

*TTS:* The TTS model converts the textual response  $\hat{\mathbf{s}}_t^{(t)}$  to its spoken form  $\hat{\mathbf{s}}_t^{(a)}$ .

This pipelined design is susceptible to error propagation, particularly from the ASR model. While each component could in principle operate directly on speech input, current speech-native models are not yet robust enough for multi-turn dialogue across languages (see supporting evidence in Table 7). We therefore adopt a pipelined implementation in this paper and release HEALTHDIAL to enable future research.

**ASR and TTS.** We evaluate a set of ASR models: whisper-1 and phi-4-MM-Inst (Aboue-

lenin et al., 2025), as well as TTS using gpt-4o-mini-tts (Achiam et al., 2023). For TTS, we condition generation on speaker demographic variables, including age group, primary language, place of origin, region of residence, and education level. Table 2 presents evaluation results for the best-performing ASR and TTS models. Additional results for the remaining models are provided in Table 6 in the Appendix. ASR models are evaluated using Word Error Rate (WER) and Character Error Rate (CER), while TTS models are assessed using Mel Cepstral Distortion (MCD) (Kubichek, 1993) and CER measured via ASR.

**Retrieval Turn Classification.** We evaluate XLM-R<sub>large</sub> (Conneau et al., 2020) fine-tuned on a training set of 500 dialogues and LLaMA3.1-8B-Inst (Grattafiori et al., 2024) with 10 randomly sampled in-context examples from the same set. Table 2 reports the best-performing model, while Table 8 provides the full results. Both models achieve over 90% accuracy, indicating the simplicity of the task, since 75.5% of dialogue turns require knowledge retrieval.

**Knowledge Selection.** In standard RAG pipelines, the retrieval model returns a candidate set of knowledge snippets, and the language model implicitly performs knowledge filtering during generation by attending to relevant content through its internal attention mechanism (Lewis et al., 2020). However, in high-stakes domains such as healthcare, interpretability is essential. To this end, we explicitly model the knowledge filtering process, following prior benchmarks that treat knowledge selection as a standalone task (Dinan et al., 2019). Accordingly, we establish three benchmarks: (i) multilingual text-to-text retrieval, (ii) multilingual speech-to-text retrieval, and (iii) explicit knowledge filtering.

We evaluate text-to-text retrieval using a set of text encoders: text-embedding-3L, gte-multilingual-B (Zhang et al., 2024), MiniLM-L12-v2 (Reimers and Gurevych, 2019), NV-Embed-v2 (Lee et al., 2025) and the statistical method BM25. We also evaluate speech-to-text retrieval using a set of multimodal multilingual encoders, including CLAP (Wu et al., 2023) and SpeechT5 (Ao et al., 2022). All models are evaluated on a parallel subset of knowledge snippets. Table 2 reports the best-performing model for each retrieval task, while Table 7 in the Appendix provides detailed results across all models. Models are evaluated using recall, precision, F1 score, and

Language	ASR		TTS		Turn Cls.	Knowledge Retrieval		Knowledge Filtering*	
	WER ↓	CER ↓	MCD ↓	CER ↓	Acc. ↑	R@10 (T) ↑	R@10 (S) ↑	EM ↑	OOK Recall ↑
Arabic	0.23	0.07	12.08	0.10	95.39	65.88	0.20	34.27	0.00
Chinese	0.24	0.14	11.46	0.17	95.23	70.63	0.23	39.19	14.29
English	0.03	0.01	11.44	0.06	96.30	75.72	0.52	44.29	42.86
Spanish	0.02	0.01	10.84	0.07	95.93	71.82	0.42	39.54	14.29
Average	0.13	0.06	11.46	0.10	95.71	71.01	0.34	39.32	17.36

Table 2: Performance of the best-performing model for each component task in HEALTHDIAL. The best model for ASR is whisper-1; for TTS, gpt-4o-mini-tts; for turn classification (Turn Cls.), XLM-R<sub>large</sub>; for text-to-text knowledge retrieval (evaluated using R@5 (T)), text-embedding-3L; for speech-to-text knowledge retrieval (evaluated using R@5 (S)), CLAP; and for knowledge filtering, gpt-4.1. (\*) For knowledge filtering, performance is reported on a randomly sampled 10% subset of the test set.

Maximal Marginal Relevance (MMR).

Based on the full results in Table 7, we observe that larger and more recent encoders, such as text-embedding-3L, consistently outperform smaller models like MiniLM-L12-v2. Additionally, multilingual text encoders exhibit notable performance disparities across languages: English achieves the highest retrieval scores, while Arabic performs the lowest, with a gap of nearly 10 points top-5 recall (R@5). Finally, we find that all evaluated multimodal encoders perform near random chance on the speech-to-text retrieval task, highlighting the difficulty of this setting and the need for future research on cross-modal alignment.

For knowledge filtering, we evaluate two approaches: (i) a threshold-based method, which retains snippets whose retrieval scores, measured by cosine similarity between the dialogue history and each candidate snippet, exceed a fixed threshold; and (ii) LLM-based methods, in which a language model is prompted to assess the relevance of each candidate snippet and retain only those it predicts relevant. We evaluate gpt-4.1-nano and LLaMA3.1-8B-Inst on the full test set, and evaluate other models from the OpenAI GPT family on 10% of the test set due to the high cost of running multilingual experiments at scale. Full results are presented in Table 8 (Appendix). Each model receives as input the top 5 retrieved snippets, selected using text-embedding-3L, and is evaluated using the Exact Match (EM) score against the ground-truth snippet set. Also, we report OOK Recall, which measures whether the model correctly returns an empty set when the query cannot be answered based on the knowledge base.

Our results highlight consistent performance disparities across languages in the knowledge selection pipeline. Despite the fully parallel experimental setup, English consistently achieves the highest retrieval and filtering accuracy, while Arabic

shows the lowest performance across all models. Furthermore, as shown in Table 3, increasing the number of retrieved candidates does not necessarily improve accuracy. While a larger candidate set raises the likelihood that the correct snippet is recalled, it also introduces more distracting snippets, which lowers filtering accuracy. This finding highlights that simply extending the input context is insufficient, and that a well-designed retrieval-augmented pipeline remains essential. In addition, Figure 3 shows that larger and more capable LLMs achieve higher performance on the deductive reasoning task of knowledge filtering, also leaving substantial room for improvement. This further demonstrates that HEALTHDIAL can serve as a multilingual benchmark for evaluating the deductive reasoning capabilities of LLMs.

The observed cross-lingual performance disparities are consistent with patterns reported in prior multilingual literature. Recent meta-analyses of multilingual benchmarks (Hu et al., 2025b) report similar language rankings, with English and Spanish typically outperforming Arabic and Chinese. Comparable trends have also been documented in multilingual ASR studies (Pratap et al., 2020; Yadav and Sitaram, 2022). Importantly, we observe consistent disparities across multiple components of the pipeline, suggesting a systematic pattern rather than a task-specific artefact.

**User Perceptions of Dialogue Systems.** We conducted a human evaluation experiment based on the Technology Acceptance Model 2 (TAM2) framework (Venkatesh and Davis, 2000) to assess user acceptance of the developed dialogue system in comparison to the WHO website. A TAM2-based questionnaire was administered to 25 participants, covering key constructs such as *Perceived Usefulness* and *Perceived Ease of Use*. All participants were fluent English speakers and did not re-

Language	Threshold	LLM-based		
		Top-5	Top-10	Top-50
Arabic	6.26	19.96	12.58	10.85
Chinese	6.61	19.86	17.15	12.28
English	6.88	23.02	23.33	18.72
Spanish	6.46	21.09	19.55	11.03
Average	6.55	21.05	18.15	13.72

Table 3: Knowledge filtering accuracy measured by Exact Match score. **Threshold** refers to a fixed similarity score used to retain relevant snippets, tuned on a validation set. **LLM@Top- $k$**  denotes filtering performed by gpt-4.1-nano over the top- $k$  retrieved snippets by text-embedding-3L. Each system turn in HEALTHDIAL is supported by at most 5 ground-truth snippets.

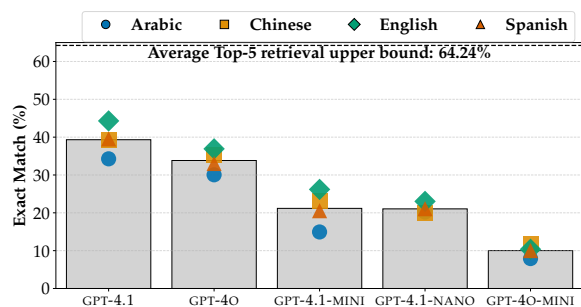


Figure 3: Knowledge filtering accuracy measured by Exact Match score for OpenAI models on 10% of the test set. Each model is tasked with selecting relevant knowledge snippets from the top 5 candidates, ranked using text-embedding-3L.

ceive any additional training. Figure 12 (Appendix) shows a screenshot of the evaluation interface, including the participant instructions, TAM2-based questionnaire, and our prototype dialogue system. Our system supports both text and speech interaction. For each system response, if available, the supporting evidence is displayed to the user.

Figure 4 presents the average ratings across key constructs. The system was implemented using gpt-4.1 as the backbone LLM. Overall, participants reported a positive attitude toward and acceptance of the system. While users found the system easy to use and generally useful, it received comparatively lower scores for perceived trustworthiness. Furthermore, the system’s overall satisfaction rating is lower than that of the WHO website. Qualitative feedback highlights the need for improvements in system output quality, user trust, and more proactive dialogue policies.

This TAM2-based evaluation is intended as an illustrative demonstration of how dialogue systems built on HEALTHDIAL can be evaluated in a standardised and reproducible manner, rather than as a

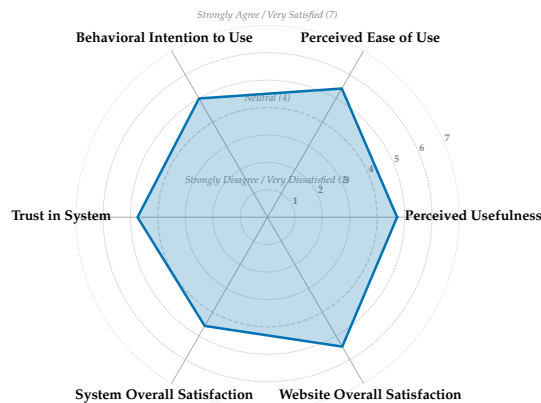


Figure 4: Average human ratings across key constructs, as reported by 25 participants. These constructs were measured using the TAM2-based questionnaire shown in Figure 12.

comprehensive assessment of cross-lingual usability or trustworthiness. Large-scale cross-lingual human evaluation remains future work.

## 5 Conclusion

We present a large-scale data collection process that produces a multilingual, multi-parallel spoken dialogue dataset for benchmarking multilingual dialogue systems. The dataset provides 6,000 dialogues and 163 hours of user speech, recorded by native speakers representing diverse language varieties across Arabic, Chinese, English, and Spanish. This dataset addresses a critical resource gap for benchmarking multilingual spoken dialogue systems and enables future evaluation of speech-native models. We benchmark a range of NLP tasks, including ASR, TTS, text-to-text and speech-to-text retrieval, and the deductive reasoning task of knowledge filtering, establishing baselines for future research.

Enabled by this dataset, future research can conduct controlled analyses of performance disparities, not only across languages, but also within language varieties and user demographics such as gender and age, and propose methods to mitigate them. Beyond standard NLP benchmarking, we release not only the dataset but also a prototype dialogue system and a complete toolkit for data collection and system evaluation. These resources will enable the research community to collect similar datasets at scale, develop spoken dialogue systems, and evaluate their performance with real-world users.

## Limitations

While HEALTHDIAL provides, to our knowledge, the first large-scale multilingual spoken dialogue benchmark, it is important to acknowledge its limitations.

**Data Collection with Synthetic Dialogues and Outline-Based Generation.** The content of HEALTHDIAL was generated with LLMs and has not been validated by healthcare professionals. We therefore release it strictly as a multilingual *language resource* for studying knowledge-grounded spoken dialogue. Although such validation could in principle be added as an extension of our pipeline, it would require substantial resources and collaboration with medical experts, which lies beyond the scope of this work. Our contribution is instead methodological: a reproducible pipeline for constructing multilingual, multi-parallel spoken dialogue datasets at scale. By combining LLMs with human annotations, we reduce both cost and privacy risks, and release a benchmark dataset that enables experiments previously not feasible with existing resources.

**Real-World Needs and Cultural Nuances.** Dialogue systems should remain responsive to real-world needs, which evolve both temporally and geographically (e.g., during the COVID-19 pandemic or in regions with a high prevalence of non-communicable diseases). Continuously retraining LLMs to reflect such changes is often infeasible (Lewis et al., 2020; Clusmann et al., 2023), motivating our use of RAG-based design. While grounding in WHO snippets ensures full parallelism across languages, it inevitably limits cultural adaptation, since the materials are not tailored to local practices. Addressing this would require collaboration with healthcare and cultural experts, which falls beyond the scope of this study. Nevertheless, by releasing a parallel benchmark, we provide a foundation for future work on culturally adapted dialogue systems.

**Cross-Lingual and Cross-Study Evaluations.** Our benchmark results reveal consistent disparities across languages, aligning with prior findings in multilingual NLP (Hu et al., 2020, 2023b; Xuan et al., 2025). However, the relative ranking of languages varies across tasks and benchmarks, reflecting a broader challenge: multilingual evaluations are often confounded by target language choice, task design, and model selection. Thus, most

benchmarks converge only on the qualitative conclusion that high-resource languages consistently outperform others. Recent work has attempted to quantify such disparities more systematically (Hu et al., 2025c), but a comprehensive cross-study comparison remains an open challenge.

**End-to-End Speech-Based Evaluation.** While HEALTHDIAL is designed to support fully speech-based system evaluation, our benchmark study follows a pipelined architecture that decomposes the system into ASR, retrieval, generation, and TTS components. This design choice reflects the current state of the field: existing speech-native language models and multimodal encoders are not yet sufficiently robust to support end-to-end spoken dialogue benchmarking at scale. In practice, the extremely low performance of current models in fully speech-based settings limits the interpretability of quantitative comparisons at present. As speech-native language models mature, HEALTHDIAL can support future evaluations of fully speech-based dialogue systems.

## Acknowledgements

This work is supported by the Cambridge–LMU Strategic Partnership grant. This work is also supported by the UK Research and Innovation (UKRI) Frontier Research Grant EP/Y031350/1 EQUATE awarded to Anna Korhonen. Songbo Hu is supported by the Cambridge International Scholarship. Ivan Vulić is supported by a Royal Society University Research Fellowship, ‘*Inclusive and Sustainable Language Technology for a Truly Multilingual World*’ (no 221137).

## References

- Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, and 1 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.
- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Junyi Ao, Rui Wang, Long Zhou, Chengyi Wang, Shuo Ren, Yu Wu, Shujie Liu, Tom Ko, Qing Li, Yu Zhang, Zhihua Wei, Yao Qian, Jinyu Li, and Furu Wei. 2022. *SpeechT5: Unified-modal encoder-decoder*

- pre-training for spoken language processing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5723–5738, Dublin, Ireland. Association for Computational Linguistics.
- Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.
- Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.
- Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations*.
- Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Rachel Bawden, Eric Bilinski, Thomas Lavergne, and Sophie Rosset. 2021. Diabla: a corpus of bilingual spontaneous written dialogues for machine translation. *Language Resources and Evaluation*, 55(3):635–660.
- Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL 2006 interactive presentation sessions*, pages 69–72.
- Isaac Caswell, Theresa Breiner, Daan van Esch, and Ankur Bapna. 2020. Language ID in the wild: Unexpected challenges on the path to a thousand-language web text corpus. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6588–6608, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- KuanChao Chu, Yi-Pei Chen, and Hideki Nakayama. 2024. Exploring and controlling diversity in llm-agent conversation. *arXiv preprint arXiv:2412.21102*.
- Jan Clusmann, Fiona R Kolbinger, Hannah Sophie Muti, Zunamys I Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P Veldhuizen, and 1 others. 2023. The future landscape of large language models in medicine. *Communications medicine*, 3(1):141.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Mark G Core and James Allen. 1997. Coding dialogs with the damsl annotation scheme. In *AAAI fall symposium on communicative action in humans and machines*, volume 56, pages 28–35. Boston, MA.
- Margaret Deuchar. 2010. Bilingual spanish-english bangor miami corpus.
- Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.
- Bosheng Ding, Junjie Hu, Lidong Bing, Mahani Aljunied, Shafiq Joty, Luo Si, and Chunyan Miao. 2022. GlobalWoZ: Globalizing MultiWoZ to develop multilingual task-oriented dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1639–1657, Dublin, Ireland. Association for Computational Linguistics.
- Haodong Duan, Jueqi Wei, Chonghua Wang, Hongwei Liu, Yixiao Fang, Songyang Zhang, Dahua Lin, and Kai Chen. 2024. BotChat: Evaluating LLMs’ capabilities of having multi-turn dialogues. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3184–3200, Mexico City, Mexico. Association for Computational Linguistics.
- Christian Federmann and William D. Lewis. 2016. Microsoft speech language translation (MSLT) corpus: The IWSLT 2016 release for English, French and German. In *Proceedings of the 13th International Conference on Spoken Language Translation*, Seattle, Washington D.C. International Workshop on Spoken Language Translation.
- Christian Federmann and William D. Lewis. 2017. The Microsoft speech language translation (MSLT) corpus for Chinese and Japanese: Conversational test data for machine translation and speech recognition. In *Proceedings of Machine Translation Summit XVI: Research Track*, pages 72–85, Nagoya Japan.
- John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Acoustics, speech, and signal processing, ieee international conference on*, volume 1, pages 517–520. IEEE Computer Society.
- Hugo Gonçalo Oliveira, Patrícia Ferreira, Daniel Martins, Catarina Silva, and Ana Alves. 2022. A brief survey of textual dialogue corpora. In *Proceedings of*

- the Thirteenth Language Resources and Evaluation Conference*, pages 1264–1274, Marseille, France. European Language Resources Association.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Yaowei Han, Sheng Li, Yang Cao, Qiang Ma, and Masatoshi Yoshikawa. 2020. **Voice-Indistinguishability: Protecting Voiceprint In Privacy-Preserving Speech Data Release**. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, Los Alamitos, CA, USA. IEEE Computer Society.
- Zhenfeng He, Yuqiang Han, Zhenqiu Ouyang, Wei Gao, Hongxu Chen, Guandong Xu, and Jian Wu. 2022. Dialmed: A dataset for dialogue-based medication recommendation. *arXiv preprint arXiv:2203.07094*.
- Charles T. Hemphill, John J. Godfrey, and George R. Doddington. 1990. **The ATIS spoken language systems pilot corpus**. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.
- Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. **XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation**. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.
- Songbo Hu, Abigail Oppong, Ebele Mogo, Charlotte Collins, Giulia Occhini, Anna Barford, and Anna Korhonen. 2025a. Natural language processing technologies for public health in africa: Scoping review. *Journal of Medical Internet Research*, 27:e68720.
- Songbo Hu, Ivan Vulić, and Anna Korhonen. 2025b. **Quantifying language disparities in multilingual large language models**. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 4003–4018, Suzhou, China. Association for Computational Linguistics.
- Songbo Hu, Ivan Vulić, and Anna Korhonen. 2025c. Quantifying language disparities in multilingual large language models. *arXiv preprint arXiv:2508.17162*.
- Songbo Hu, Han Zhou, Mete Hergul, Milan Gritta, Guchun Zhang, Ignacio Iacobacci, Ivan Vulić, and Anna Korhonen. 2023a. **Multi 3 WOZ: A multilingual, multi-domain, multi-parallel dataset for training and evaluating culturally adapted task-oriented dialog systems**. *Transactions of the Association for Computational Linguistics*, 11:1396–1415.
- Songbo Hu, Han Zhou, Moy Yuan, Milan Gritta, Guchun Zhang, Ignacio Iacobacci, Anna Korhonen, and Ivan Vulić. 2023b. **A systematic study of performance disparities in multilingual task-oriented dialogue systems**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6825–6851, Singapore. Association for Computational Linguistics.
- Chia-Chien Hung, Anne Lauscher, Ivan Vulić, Simone Ponzetto, and Goran Glavaš. 2022. **Multi2WOZ: A robust multilingual dataset and conversational pre-training for task-oriented dialog**. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3687–3703, Seattle, United States. Association for Computational Linguistics.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.
- Ye Jia, Michelle Tadmor Ramanovich, Quan Wang, and Heiga Zen. 2022. **CVSS corpus and massively multilingual speech-to-speech translation**. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6691–6703, Marseille, France. European Language Resources Association.
- Aditya Joshi, Raj Dabre, Diptesh Kanojia, Zhuang Li, Haolan Zhan, Gholamreza Haffari, and Doris Dipold. 2025. Natural language processing for dialects of a language: A survey. *ACM Computing Surveys*, 57(6):1–37.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. **The state and fate of linguistic diversity and inclusion in the NLP world**. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.
- Seokhwan Kim, Luis Fernando D’Haro, Rafael E. Banchs, Jason D. Williams, Matthew Henderson, and Koichiro Yoshino. 2016. **The fifth dialog state tracking challenge**. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 511–517.
- Seokhwan Kim, Yang Liu, Di Jin, Alexandros Papanagelis, Karthik Gopalakrishnan, Behnam Hedayatnia, and Dilek Hakkani-Tür. 2021. “how robust ru?”: Evaluating task-oriented dialogue systems on spoken conversations. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1147–1154. IEEE.
- R. Kubichek. 1993. **Mel-cepstral distance measure for objective speech quality assessment**. In *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, volume 1, pages 125–128 vol.1.

- Khai Le-Duc, Tuyen Tran, Bach Phan Tat, Nguyen Kim Hai Bui, Quan Dang, Hung-Phong Tran, Thanh-Thuy Nguyen, Ly Nguyen, Tuan-Minh Phan, Thi Thu Phuong Tran, and 1 others. 2025. Multimedst: Large-scale many-to-many multilingual medical speech translation. *arXiv preprint arXiv:2504.03546*.
- Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. [NV-embed: Improved techniques for training LLMs as generalist embedding models](#). In *The Thirteenth International Conference on Learning Representations*.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.
- Dongdong Li, Zhaochun Ren, Pengjie Ren, Zhumin Chen, Miao Fan, Jun Ma, and Maarten de Rijke. 2021. [Semi-supervised variational reasoning for medical dialogue generation](#). In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '21*, page 544–554, New York, NY, USA. Association for Computing Machinery.
- 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.
- Zhaojiang Lin, Zihan Liu, Genta Indra Winata, Samuel Cahyawijaya, Andrea Madotto, Yejin Bang, Etsuko Ishii, and Pascale Fung. 2021a. [XPersona: Evaluating multilingual personalized chatbot](#). In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, pages 102–112, Online. Association for Computational Linguistics.
- Zhaojiang Lin, Andrea Madotto, Genta Indra Winata, Peng Xu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale Fung. 2021b. [BiToD: A bilingual multi-domain dataset for task-oriented dialogue modeling](#). In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks 1, NeurIPS Datasets and Benchmarks 2021, December 2021, virtual*.
- Wenge Liu, Jianheng Tang, Yi Cheng, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. 2022. Meddg: an entity-centric medical consultation dataset for entity-aware medical dialogue generation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 447–459. Springer.
- Yinhong Liu, Yimai Fang, David Vandyke, and Nigel Collier. 2024a. [TOAD: Task-oriented automatic dialogs with diverse response styles](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8341–8356, Bangkok, Thailand. Association for Computational Linguistics.
- Yinhong Liu, Jianfeng He, Hang Su, Ruixue Lian, Yi Nian, Jake Vincent, Srikanth Vishnubhotla, Robinson Piramuthu, and Saab Mansour. 2025. Mdseval: A meta-evaluation benchmark for multimodal dialogue summarization. *arXiv preprint arXiv:2510.01659*.
- Yinhong Liu, Yixuan Su, Ehsan Shareghi, and Nigel Collier. 2024b. [Unlocking structure measuring: Introducing PDD, an automatic metric for positional discourse coherence](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 92–100, Mexico City, Mexico. Association for Computational Linguistics.
- Zeming Liu, Ping Nie, Jie Cai, Haifeng Wang, Zheng-Yu Niu, Peng Zhang, Mrinmaya Sachan, and Kaiping Peng. 2023. [XDailyDialog: A multilingual parallel dialogue corpus](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12240–12253, Toronto, Canada. Association for Computational Linguistics.
- Olga Majewska, Evgeniia Razumovskaia, Edoardo Maria Ponti, Ivan Vulić, and Anna Korhonen. 2023. [Cross-lingual dialogue dataset creation via outline-based generation](#). *Transactions of the Association for Computational Linguistics*, 11:139–156.
- Silvano Martello and Paolo Toth. 1987. Linear assignment problems. In *North-Holland Mathematics Studies*, volume 132, pages 259–282. Elsevier.
- Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. [Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints](#). *Transactions of the Association for Computational Linguistics*, 5:309–324.
- Ossama Obeid, Nasser Zalmout, Salam Khalifa, Dima Taji, Mai Oudah, Bashar Alhafni, Go Inoue, Fadhl Eryani, Alexander Erdmann, and Nizar Habash. 2020. [CAMEL tools: An open source python toolkit for Arabic natural language processing](#). In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7022–7032, Marseille, France. European Language Resources Association.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [Bleu: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Edoardo Maria Ponti, Helen O’Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. [Modeling](#)

- language variation and universals: A survey on typological linguistics for natural language processing. *Computational Linguistics*, 45(3):559–601.
- Matt Post, Gaurav Kumar, Adam Lopez, Damianos Karakos, Chris Callison-Burch, and Sanjeev Khudanpur. 2013. Improved speech-to-text translation with the fisher and callhome Spanish-English speech translation corpus. In *Proceedings of the 10th International Workshop on Spoken Language Translation: Papers*, Heidelberg, Germany.
- Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MIs: A large-scale multilingual dataset for speech research. In *Interspeech*.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pages 28492–28518. PMLR.
- Vikram Ramanarayanan and David Suendermann-Oeft. 2017. Jee haan, i'd like both, por favor: Elicitation of a code-switched corpus of hindi-english and spanish-english human-machine dialog. In *Interspeech*, pages 47–51.
- Evgeniia Razumovskaia, Goran Glavaš, Olga Majewska, Edoardo Maria Ponti, Anna Korhonen, and Ivan Vulić. 2022. Crossing the conversational chasm: A primer on natural language processing for multilingual task-oriented dialogue systems. *Journal of Artificial Intelligence Research*, 74:1351–1402.
- 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 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Stephen Robertson, Hugo Zaragoza, and 1 others. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.
- Jorge A Rodriguez, Emily Alsentzer, and David W Bates. 2024. Leveraging large language models to foster equity in healthcare. *Journal of the American Medical Informatics Association*, 31(9):2147–2150.
- Anna Rogers, Timothy Baldwin, and Kobi Leins. 2021. ‘just what do you think you’re doing, dave?’ a checklist for responsible data use in NLP. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4821–4833, Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215.
- Ali Shahin Shamsabadi, Brij Srivastava, Aurélien Bellet, Nathalie Vauquier, Emmanuel Vincent, Mohamed Maouche, Marc Tommasi, and Nicolas Papernot. 2023. Differentially private speaker anonymization. *Proceedings on Privacy Enhancing Technologies*, 2023:98–114.
- Shuichiro Shimizu, Chenhui Chu, Sheng Li, and Sadao Kurohashi. 2023. Towards speech dialogue translation mediating speakers of different languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1122–1134, Toronto, Canada. Association for Computational Linguistics.
- Shuzheng Si, Wentao Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. 2023. Spokenwoz: a large-scale speech-text benchmark for spoken task-oriented dialogue agents. In *Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23*, Red Hook, NY, USA. Curran Associates Inc.
- Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.
- Jamal Uddin, Cheng Feng, and Junfang Xu. 2025. Health communication on the internet: Promoting public health and exploring disparities in the generative ai era. *Journal of Medical Internet Research*, 27:e66032.
- Viswanath Venkatesh and Fred D Davis. 2000. A theoretical extension of the technology acceptance model: Four longitudinal field studies. *Management science*, 46(2):186–204.
- World Bank. 2023. Digital-in-Health: Unlocking the Value for Everyone. <https://openknowledge.worldbank.org/handle/10986/40212>. License: CC BY 4.0.
- Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.
- Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Yun Xing, Junjue Wang, Huitao Li, Xin Li, Kunyu Yu, and 1 others. 2025. Mmlu-prox: A multilingual benchmark for advanced large language model evaluation. *arXiv preprint arXiv:2503.10497*.
- Hemant Yadav and Sunayana Sitaram. 2022. A survey of multilingual models for automatic speech recognition. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5071–5079, Marseille, France. European Language Resources Association.

Steve Young, Milica Gašić, Blaise Thomson, and Jason D. Williams. 2013. [Pomdp-based statistical spoken dialog systems: A review](#). *Proceedings of the IEEE*, 101(5):1160–1179.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. [MedDialog: Large-scale medical dialogue datasets](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, and 1 others. 2024. [mgte: Generalized long-context text representation and reranking models for multilingual text retrieval](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412.

Lei Zuo, Kun Qian, Bowen Yang, and Zhou Yu. 2021. [AllWOZ: Towards multilingual task-oriented dialog systems for all](#). *CoRR*, abs/2112.08333. Version 1.

## A Supplementary Details for Replication

We provide supplementary information to support the replication of the data creation process and benchmark experiments described in this paper. We disclose the use of AI assistants for code writing and editorial assistance during the preparation of this work.

**Figure 5** illustrates the intended use case of dialogue systems developed using HEALTHDIAL, highlighting the integration of grounding in external health knowledge.



Figure 5: An illustration of dialogue systems based on HEALTHDIAL. Each dialogue turn is explicitly grounded in external knowledge snippets sourced from the WHO website. While the illustration shows a text-based dialogue in English, HEALTHDIAL supports both text and speech in Arabic, Chinese, English, and Spanish.

### A.1 Knowledge Base Construction

**Example 1** shows a sample knowledge snippet collected in English in HEALTHDIAL.

Example 1: Example of a health knowledge snippet in JSON format.

```
{
  "url": "https://www.who.int/news-room/questions-and-answers/item/traditional-medicine",
  "language": "ENG",
  "data": {
    "type": "qa_pair",
    "topic": "Traditional medicine",
    "title": "What is traditional medicine?",
    "content": "Traditional medicine has a long history. It is the sum of the knowledge, skills and practices based on the theories, beliefs and experiences indigenous to different cultures, whether explicable or not, used in the maintenance of health and the prevention, diagnosis, improvement or treatment of physical and mental illness."
  },
  "parallel_data": true,
  "parallel_identifier": "questions-and-answers/item/traditional-medicine::0",
  "unique_identifier": "6a85e2b5-ee53-493f-82ca-26488110b593"
}
```

**Knowledge Snippet Collection.** We crawled webpages from the WHO website on 11 May 2025, as described in §3. As shown in Figure 6, each collapsible panel corresponds to a single knowledge snippet, with the page heading serving as the topic, the panel title as the title, and the panel content as the content of the snippet. Multiple snippets can be derived from the same page, all sharing the same topic. Overall, we extracted 4,785 English snippets ( $\mathbb{K}^{\text{ENG}}$ ) from 590 webpages, 2,317 Arabic snippets ( $\mathbb{K}^{\text{ARA}}$ ) from 299 webpages, 2,431 Chinese snippets ( $\mathbb{K}^{\text{ZHO}}$ ) from 303 webpages, and 2,512 Spanish snippets ( $\mathbb{K}^{\text{SPA}}$ ) from 307 webpages.

**Knowledge Snippet Alignment.** Each WHO health topic page typically contains multiple knowledge snippets, which may not always appear in the same order or have exact one-to-one correspondence across different languages. In order to construct four parallel sets of knowledge snippets, we identify the largest possible set of matched snippets that are aligned across all four languages for each page. We approach this problem by assuming that the English knowledge snippets form a superset of the others; that is,  $\mathbb{K}^{\text{ARA}}, \mathbb{K}^{\text{ZHO}}, \mathbb{K}^{\text{SPA}} \subseteq \mathbb{K}^{\text{ENG}}$ . The task is then transformed into assigning each snippet in the other three languages to a correspond-

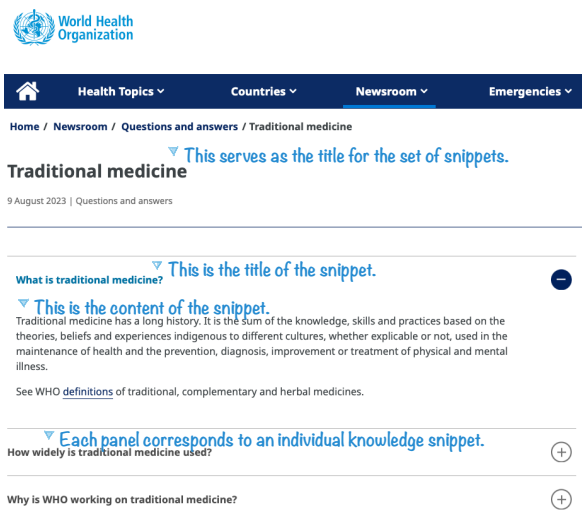


Figure 6: A screenshot of a WHO webpage. The figure is annotated to show how each component of the webpage corresponds to the attributes of a knowledge snippet.

ing snippet in English. We define the alignment as a set of functions that map each non-English snippet to a corresponding English snippet:  $f^{\text{LAN}} : \mathbb{K}^{\text{LAN}} \rightarrow \mathbb{K}^{\text{ENG}}$  for  $\text{LAN} \in \{\text{ARA}, \text{ZHO}, \text{SPA}\}$ , where  $f^{\text{LAN}}(\mathbf{k})$  returns the English snippet in  $\mathbb{K}^{\text{ENG}}$  that is semantically equivalent to snippet  $\mathbf{k} \in \mathbb{K}^{\text{LAN}}$ . This alignment problem can be modelled as a *linear sum assignment problem* (Martello and Toth, 1987), which seeks an optimal assignment of ‘tasks’ to ‘workers’ that minimises the total cost. In our case, the cost is defined as the semantic distance between snippets in English and those in the other three languages. We compute pairwise semantic distance as  $1 - \text{cosine similarity}$  between snippet embeddings, where each snippet is represented by the concatenation of its **title** and **content**. Embeddings are generated using OpenAI’s text-embedding-3-L model. The optimal assignments are computed using the *Hungarian algorithm* (Martello and Toth, 1987).

## A.2 Pilot Experiments

**Example 2** shows one example scenario used in the pilot experiment. It was generated by ChatGPT and verified by the research team.

### Example 2: Burn injury scenario

You recently experienced a burn injury while cooking at home. The affected area is painful and appears red and swollen. You’re unsure

about the severity of the burn and whether you should seek medical attention. Questions arise about how to properly care for the burn at home, what signs indicate a need for professional medical help, and how long it will take to heal. These concerns are important to address promptly to ensure proper treatment and avoid complications.

**Example 3** shows the system prompt used to develop the pilot system. The model checkpoint is gpt-4o, with a temperature of 0.5 and a top\_p of 0.9 used for the sampling method.

### Example 3: Pilot system prompt

You are a health advisor and please try to answer the following question from a patient. You should provide a brief response to the patient’s question. Your response should also be coherent with the dialog history. Users might not always have access to medical professionals. Please try your best to answer their questions Please only output the response but nothing else.

**Example 4** presents the 11 dialogue acts we constructed, each with an example utterance, based on the 20 pilot dialogues described in §3.

### Example 4: Dialogue Acts in HEALTHDIAL

**1. Opening:** The *system* initiates the conversation with a greeting and an introduction to its role or the service provided.

*Example (System):* Hello, I’m your virtual health assistant. How can I help you today?

**2. Health Concern Presentation:** The *user* states their primary health concern, symptom, or question.

*Example (User):* Hey, I burned my hand cooking last week. It’s really painful, red, and swollen.

**3. Information Gathering:** The *system* asks clarification questions to gather more context about the user’s symptoms or medical history.

*Example (System):* Were you vaccinated for yellow fever before your trip?

**4. Explanation / Medical Education:** The *system* provides in-depth information or educates the user about their condition, treatment

options, and preventive measures.

*Example (System):* If the burn is larger than 3 inches or on your face, hands, or joints, you should definitely see a doctor.

**5. Care Planning and Guidance:** The *system* offers specific advice on managing the health issue, including treatment options, preventive measures, lifestyle modifications, and self-care techniques.

*Example (System):* Until you see a doctor, keep the burn clean and covered with a sterile, non-stick bandage.

**6. Decision Support:** The *user or system* may discuss different options, relevant risks and benefits, and explore user preferences.

*Example (System):* It's important to consider your options and what feels right for you. You can also seek support from a trusted friend, family member, or a professional counsellor.

**7. Healthcare System Navigation:** The *user or system* may discuss guidance on navigating the healthcare system, including finding a provider, making an appointment, and understanding insurance coverage and costs.

*Example (System):* You can find a local urgent care centre or call your primary care doctor to schedule an appointment.

**8. Legal and Ethical Considerations:** The *user or system* may discuss legal and ethical considerations, including informed consent and patient rights.

*Example (System):* In the UK, your medical records are confidential and protected by law.

**9. Privacy and Confidentiality:** The *user or system* may inquire about, or proactively assure, the privacy and confidentiality of the user's information.

*Example (System):* Your information is safe with us. We take your privacy very seriously.

**10. Emotional Support:** The *system* offers emotional support, empathy, and reassurance to the user. *Example (System):* I'm sorry to hear that you're going through this. It's completely normal to feel scared and overwhelmed.

**11. Closing:** The *system* ends the conversation with a summary, an offer of further assistance, or a farewell. *Example (System):* You're welcome! Take care, and I hope you feel better soon. Goodbye!

### A.3 Dialogue Schemata Creation

**Figure 7** shows the transition probabilities in our hierarchical Markov model. Let  $\mathbf{a}_i^u$  and  $\mathbf{a}_i^s$  denote the discourse acts associated with the  $i$ -th user and system turns, respectively, and let  $\mathbf{a}_0^s$  denote the initial system act, which is fixed across all dialogues (i.e., an *Opening* act introducing the system). The full *dialogue schema* is defined as:  $\mathcal{A} = [\mathbf{a}_0^s, \mathbf{a}_1^u, \mathbf{a}_1^s, \dots, \mathbf{a}_n^u, \mathbf{a}_n^s]$ . We factorise the probability of the schema (excluding the fixed first act) as:  $P(\mathcal{A}) = \prod_{i=1}^n P(\mathbf{a}_i^u | \mathbf{a}_{i-1}^u) \cdot P(\mathbf{a}_i^s | \mathbf{a}_i^u)$ , where  $P(\mathbf{a}_i^u | \mathbf{a}_{i-1}^u)$  represents the user-to-user transition probabilities, and  $P(\mathbf{a}_i^s | \mathbf{a}_i^u)$  models the system's response act conditioned on the current user act.

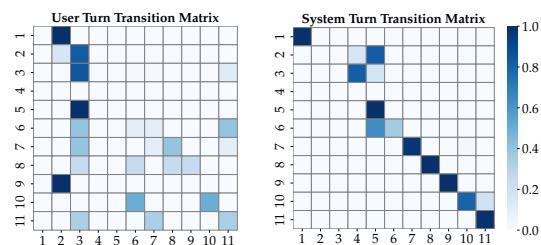


Figure 7: Transition probabilities in our Markov model. The left plot shows user turn transitions, namely  $P(\mathbf{a}_i^u | \mathbf{a}_{i-1}^u)$ , while the right plot shows system turn transitions,  $P(\mathbf{a}_i^s | \mathbf{a}_i^u)$ . Role indices correspond to the discourse role schema presented in Example 4.

### A.4 Surface Realisation.

**Figure 8** shows a screenshot of the annotation interface with the guidelines shown to annotators. Annotators for Arabic, English, and Spanish were presented with guidelines in English, except for Chinese annotators, who received the guidelines in Chinese. This exception was made possible due to the availability of native Chinese-speaking researchers on our team who translated the website.

### A.5 Examples and Statistics of HEALTHDIAL

**Figure 9** presents a set of parallel dialogues in four languages, English, Arabic, Chinese, and Spanish, drawn from the HEALTHDIAL dataset.

**Figure 10** and **Figure 11** show the distribution of dialogues by annotator gender and age group, respectively, for each language.

**Table 4** reports word-level and subword-level statistics across the four languages. Across all languages, human-authored user utterances are consistently longer and more lexically diverse than those gener-

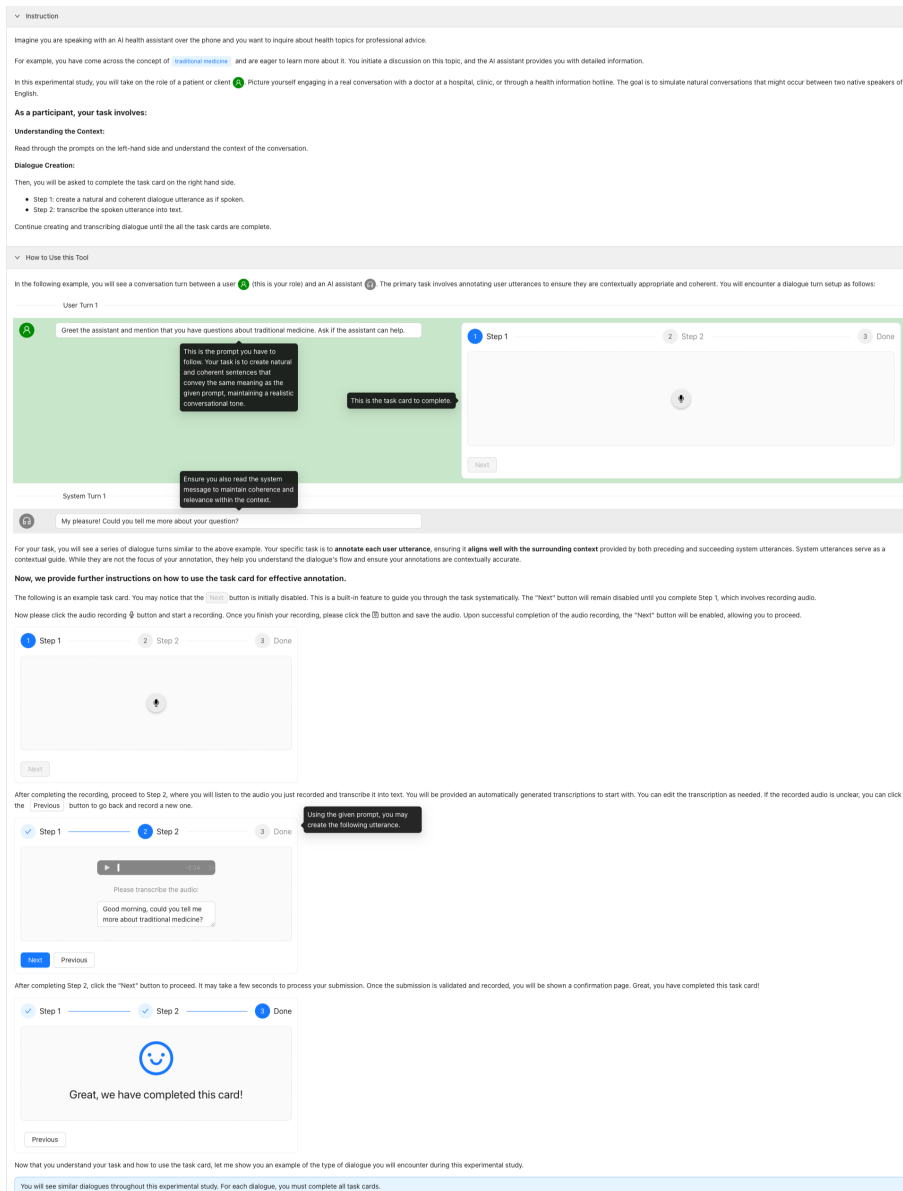


Figure 8: Screenshot of the annotation interface with the guidelines shown to annotators during English data collection.

ated by LLMs. Dialogues contain an average of 6.5 user turns. In English, for example, human utterances contain 35.71 tokens on average, compared to 18.66 tokens for LLM-generated counterparts, with a substantially larger vocabulary size. Similar patterns are observed in Arabic, Chinese, and Spanish. These results indicate that the outline-based data collection methodology effectively elicits more diverse and naturalistic user utterances than direct LLM generation.

## A.6 Experimental Setup

**Table 5** lists all the language models we used in this work, along with their respective checkpoints in the Huggingface repository and the OpenAI API.

**Figure 12** shows a screenshot of the human evaluation interface, including the guidelines provided to annotators. Questions 1 and 3 measure *Perceived Usefulness*, Question 2 measures *Perceived Ease of Use*, Question 4 measures *Behavioural Intention to Use*, Question 7 measures *Trust in System*, Question 5 measures *Overall Satisfaction with the Dialogue System*, and Question 10 measures *Overall Satisfaction with the WHO website*. The image also illustrates the user interface of our prototype dialogue system, which can be either embedded in a webpage or deployed as a stand-alone application. The system supports both text and speech interactions.

<b>System Turn S0:</b>	Good day! I'm your AI healthcare assistant. How may I assist you today?	مرحبًا! أنا مساعدك الصحي الذكي. كيف يمكنني مساعدتك اليوم؟	你好！我是你的AI医疗助手。今天我能为你做些什么？	¡Buenos días! Soy tu asistente de salud AI. ¿Cómo puedo ayudarte hoy?
<b>Original LLM-Generated User Turn U1:</b>	Hi there. I've been hearing a lot about new COVID-19 vaccines. Are they safe to use?	مرحبًا. سمعت الكثير عن لقاحات جديدة لفيروس كورونا. هل هي آمنة للاستخدام؟	你好。我最近听说了很多关于新冠疫苗的消息。它们安全吗？	Hola. He estado escuchando mucho sobre las nuevas vacunas contra el COVID-19. ¿Son seguras para usar?
<b>Human Created User Turn U1:</b>	Good day to you too. I have been hearing so much about the new covid 19 vaccines, and I am really curious to learn more. I am rather concerned about their safety and hope you can give me some more information on this topic.	أفهم قلقك بشأن سلامة لقاحات فيروس كورونا الجديدة. لدى منظمة الصحة العالمية إجراء يسمى قائمة الاستخدام الطارئ (EUL) لتقييم اللقاحات الجديدة خلال حالات الطوارئ (EUL) الصحية العامة. هذا الإجراء يقيم جودة وسلامة وفعالية اللقاحات. حتى لو كانت لا تزال قيد التطوير.	您好,小助手,我最近看到了许多关于新冠疫苗的新闻,正面的,负面的都有,我对新冠疫苗的安全性和有效性特别好奇,也很担心,请问您有关于新冠疫苗更多的信息吗?	Hola, buenos días. Me gustaría saber algo más sobre las vacunas nuevas que hay del COVID. Lo he escuchado en las noticias y me gustaría saber si son seguras.
<b>System Turn S1:</b>	I understand your concern about the safety of new COVID-19 vaccines. The World Health Organization (WHO) has a procedure called Emergency Use Listing (EUL) to assess new vaccines during public health emergencies. This process evaluates the quality, safety, and efficacy of vaccines, even if they're still in development.	أفهم قلقك بشأن سلامة لقاحات فيروس كورونا الجديدة. لدى منظمة الصحة العالمية إجراء يسمى قائمة الاستخدام الطارئ (EUL) لتقييم اللقاحات الجديدة خلال حالات الطوارئ (EUL) الصحية العامة. هذا الإجراء يقيم جودة وسلامة وفعالية اللقاحات. حتى لو كانت لا تزال قيد التطوير.	我理解你对新冠疫苗安全性的担忧。世界卫生组织 (WHO) 有一个叫做紧急使用清单 (EUL) 的程序,用于在公共卫生紧急情况下评估新疫苗。这个过程会评估疫苗的质量、安全性和有效性,即使它们仍在开发中。	Entiendo tu preocupación sobre la seguridad de las nuevas vacunas contra el COVID-19. La Organización Mundial de la Salud (OMS) tiene un procedimiento llamado Lista de Uso de Emergencia (EUL) para evaluar nuevas vacunas durante emergencias de salud pública. Este proceso evalúa la calidad, seguridad y eficacia de las vacunas, incluso si todavía están en desarrollo.
<b>Supporting Knowledge Snippet K1:</b>	<b>Coronavirus disease (COVID-19) - When can the EUL be used?</b> The EUL is used during public health emergencies. When products are not licensed yet (still in development), WHO will assess the quality, safety and efficacy (or performance) data generated during development and conduct a risk-benefit assessment to decide if they can be used outside clinical trials.	<b>مرض فيروس كورونا (كوفيد-19) - متى يمكن استخدام بروتوكول الاستعمالات الطارئة؟</b> يستخدم بروتوكول الاستعمالات الطارئة أثناء الطوارئ الصحية العامة. ومتى كانت المنتجات غير مُرخَّصة بعد (لا تزال قيد التطوير), ستقوم المنظمة بتقييم بيانات الجودة والمأمونية والفعالية (أو الأداء) التي تولَّد أثناء التطوير وإجراء تقييم للمخاطر مقابل الفوائد للبت في إمكان استخدامها خارج التجارب السريرية.	<b>2019冠状病毒病 (COVID-19) - 何时可以利用紧急使用列表程序?</b> 紧急使用列表程序应在突发公共卫生事件期间使用。当产品尚未获得许可 (仍在开发中) 时,世卫组织将评估开发过程中产生的质量、安全性和效力 (性能) 数据,并进行风险效益评估,以决定有关产品是否可以在临床试验之外使用。	<b>Enfermedad por el coronavirus de 2019 (COVID-19) - ¿Cuándo puede utilizarse el EUL?</b> El EUL se utiliza durante las emergencias de salud pública. Cuando los productos aún no han sido autorizados (todavía están en fase de desarrollo), la OMS evalúa los datos de calidad, seguridad y eficacia (o desempeño) generados durante el desarrollo y lleva a cabo una evaluación de riesgos y beneficios para decidir si pueden utilizarse fuera de los ensayos clínicos.

Figure 9: Example set of parallel dialogues in four languages, English, Arabic, Chinese, and Spanish, extracted from the HEALTHDIAL dataset. Due to space limitations, we show only the first three turns of each dialogue. For each user turn, both the LLM-generated and the human-produced utterances are provided. As shown, human-authored utterances tend to be more complex and conversational than those generated by the LLM. The dialogue ID for this example is LAN\_12.

Language	User Utterances (Word)			Generated Utterances (Word)			User Utterances (LLaMA Subword)		
	# Tokens	# Words	TTR	# Tokens	# Words	TTR	# Tokens	# Words	TTR
Arabic	356,673	24,165	0.068	154,832	7,374	0.048	695,958	2,477	0.0036
English	348,042	6,327	0.018	181,838	3,586	0.020	355,782	6,846	0.0192
Spanish	369,805	10,141	0.027	167,672	5,396	0.032	507,561	5,275	0.0104
Chinese	293,954	11,153	0.038	165,107	4,161	0.025	376,343	3,723	0.0099
Average	342,619	12,947	0.038	167,862	5,129	0.031	483,661	4,580	0.0108

Table 4: Comparison of word-level and subword-level statistics across four languages. **User Utterances** correspond to human-annotated transcriptions provided by annotators, while **Generated Utterances** correspond to LLM-generated responses. **# Tokens** denotes the total number of tokens across all utterances. **# Words** refers to the number of unique tokens (i.e., vocabulary size). **TTR** represents the type-token ratio, measuring lexical diversity within each language. Tokenisation is performed at the word level using CAMEL Tools (Obeid et al., 2020) for Arabic, jieba (github.com/fxsjy/jieba) for Chinese, and NLTK (Bird, 2006) for English and Spanish, while subword-level analysis relies on the unified LLaMA3.1-8B-Inst tokeniser.

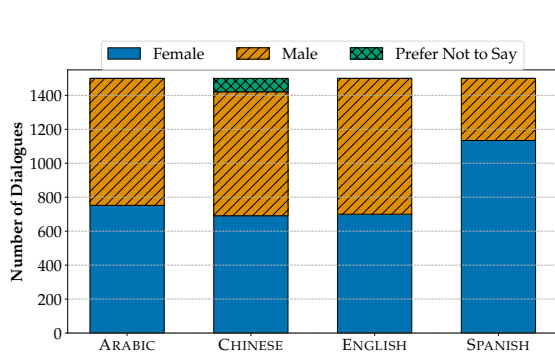


Figure 10: Distribution of dialogues by annotator gender for each language in HEALTHDIAL.

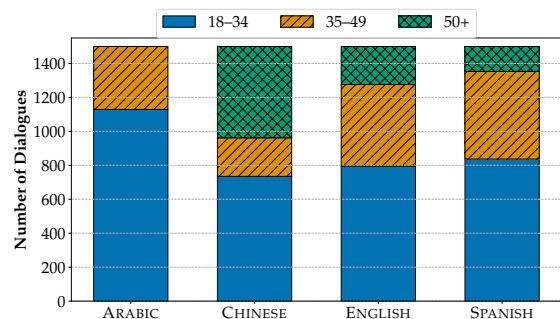


Figure 11: Distribution of dialogues by annotator age group for each language in HEALTHDIAL. The original data includes finer-grained age groups, which we cluster into broader categories for visualisation.

Model	Checkpoint
<b>Huggingface</b>	
whisper-L-v3	openai/whisper-large-v3
phi-4-MM-Inst	microsoft/Phi-4-multimodal-instruct
XLNet <sub>large</sub>	xlm-roberta-large
LlAMA3.1-8B-Inst	meta-llama/Llama-3.1-8B-Instruct
gte-multilingual-B	Alibaba-NLP/gte-multilingual-base
MiniLM-L12-v2	sentence-transformers/all-MiniLM-L12-v2
NV-Embed-v2	nvidia/NV-Embed-v2
SpeechT5	microsoft/speecht5_asr
<b>OpenAI</b>	
gpt-4.1	gpt-4o-2024-05-13
gpt-4.1-mini	gpt-4.1-mini-2025-04-14
gpt-4.1-nano	gpt-4.1-nano-2025-04-14
gpt-4o (data construction)	gpt-4o-2024-05-13
gpt-4o (benchmark)	gpt-4o-2024-11-20
gpt-4o-mini	gpt-4o-mini-2024-07-18
whisper-1	whisper-1
gpt-4o-mini-tts	gpt-4o-mini-tts
text-embedding-3L	text-embedding-3-large

Table 5: Language models used in our experiments, along with their corresponding HuggingFace or OpenAI checkpoints. Note that the gpt-4o model used for benchmarking is a more recent version than the one used during data construction. For the CLAP model, we use the 630k-audioset-best.pt checkpoint from its official GitHub repository: [github.com/LAION-AI/CLAP](https://github.com/LAION-AI/CLAP).

Step 1: Instruction

In this task, you will explore two different tools for finding health-related information:

- The official WHO Website
- A Dialogue System that allows you to ask health-related questions through conversation

First, we will ask you to visit the WHO website to search for information. After that, you will talk with our assistant by typing your questions or requests in a chat-like format. You will be given a hypothetical health-related scenario (e.g., back pain, vaccine concerns, or a skin condition). Please try to find useful and trustworthy information on both platforms about your assigned scenario.

- You've been experiencing changes in your mood and behavior lately that are concerning you.
- Sometimes you feel unusually low and find it hard to get out of bed, while at other times, you feel irritable and restless for no apparent reason.
- You've also noticed changes in your sleep patterns and appetite.
- These symptoms have been affecting your daily life and relationships, and you're starting to wonder if they could be signs of a mental health issue.
- You know it's essential to address these concerns, but you're not sure where to start.
- You have questions about what these symptoms might indicate, whether they could be linked to a specific mental disorder, and what steps you should take next.
- It's normal to have these questions and important to seek answers so that you can understand what's going on and get the support you need.

Imagine it's a real situation you or someone you care about is facing. Search naturally — just as you would if you were genuinely trying to solve a health concern. At the end of the task, you will be asked to fill out a short questionnaire comparing your experience with both systems. There are no right or wrong answers — we are simply interested in your opinion.

Step 2: Visiting WHO Website

Please visit the official WHO website to look for health information related to your assigned scenario. Try to find useful and trustworthy answers — just as you would in a real-life situation. You can browse articles, use the search bar, and explore the relevant pages.

When you're finished exploring, return to this page to move on to the chatbot.


[Open WHO Website in New Tab](#)


Step 3: Chat with Chatbot

Now it's time to chat with our health assistant chatbot! You can ask questions, explain your situation, and follow up naturally — just like talking to a doctor or online support service. The chatbot supports both text and voice. You can type your questions, or speak directly by clicking the microphone icon in the bottom-right corner. Try to use the chatbot to find the same kind of information you searched for on the WHO website. Feel free to ask in your own words and follow up naturally if something is unclear. After you finish the conversation, you'll be asked to complete a short questionnaire to share your experience.

**WHO Health Advice Chatbot**

System connected. You are now chatting with the WHO health-advisor chatbot.

 Hello, I'm your WHO health-advisor. How can I help you today?

 33"

[Hide Evidence](#)

**What are the signs of stress?**

Stress makes it hard for us to relax and can come with a range of emotions, including anxiety and irritability. When stressed, we may find it difficult to concentrate. We may experience headaches or other body pains, an upset stomach or trouble sleeping. We may find we lose our appetite or eat more than usual. Chronic stress can worsen pre-existing health problems and may increase our use of alcohol, tobacco and other substances. Stressful situations can also cause or exacerbate mental health conditions, most commonly anxiety and depression, which require access to health care. When we suffer from a mental health condition, it may be because our symptoms of stress have become persistent and have started affecting our daily functioning, including at work or school.

<https://www.who.int/news-room/questions-and-answers/item/stress>

Type your message here...

Step 4: Evaluate and Compare

Please evaluate your experience using the dialogue system compared to the WHO website. There are no right or wrong answers — we're simply interested in your honest impressions. Your feedback will help us improve the system.

For the following questions, please rate from 1 to 7:  
**1 = Strongly Disagree    4 = Neutral    7 = Strongly Agree**

- The dialogue system helped me understand health information better than the WHO website.
- The dialogue system was easier to use than the WHO website.
- The dialogue system gave me more useful answers than the WHO website.
- I would prefer using the dialogue system over the WHO website in the future.
- Overall satisfaction with the dialogue system?
- Did the system help you find the answers?  
 Yes    Partially    No
- I trust the information provided by the dialogue system more than the WHO website.
- If you had a health concern, which tool would you use first?  
 Dialogue System    WHO Website    It depends
- Why did you choose that option?
- Overall, how satisfied were you with the WHO website?
- How much of the needed information did you find?  
 All of it    Most of it    Some of it    Very little or none
- Any additional feedback or suggestions?

Figure 12: Screenshot of the human evaluation interface with guidelines shown to annotators. The screenshot also illustrates the user interface of our prototype dialogue system, which can be embedded within a webpage or used as a stand-alone application. The system supports both text and speech interaction. For each system response, if available, the corresponding supporting evidence can be displayed to the user.

## B Additional Results for Benchmarking

This section presents experimental results that complement the main retrieval benchmarks discussed in §4.

**Table 6** reports the full evaluation results of the ASR and TTS models. WER assumes word-level tokenisation based on white-space, which is not directly applicable to Chinese. To address this, we pre-tokenise Chinese transcriptions using the *jieba* segmentation tool.

**Table 7** shows the complete evaluation results for both text-to-text and speech-to-text retrieval tasks.

**Table 8** presents evaluation results for retrieval turn classification and knowledge filtering across four languages.

**Table 9** presents the complete evaluation results for response generation, using BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) as evaluation metrics. However, traditional reference-based metrics offer limited insight into the actual quality and utility of system responses from the perspective of end users. Moreover, a key limitation of this work is that the system responses have not been validated by healthcare professionals. As such, HEALTHDIAL should not be used as a ground-truth reference for evaluating response generation models in the health domain.

Language	Model	WER ↓	CER ↓	MCD ↓	CER via ASR ↓	Task
Arabic	whisper-1	0.23	0.07	—	—	ASR
	phi-4	5.89	5.79	—	—	ASR
	gpt-4o-mini-tts	—	—	12.08	0.10	TTS
Chinese	whisper-1	0.24	0.14	—	—	ASR
	phi-4	1.03	0.78	—	—	ASR
	gpt-4o-mini-tts	—	—	11.46	0.17	TTS
English	whisper-1	0.03	0.01	—	—	ASR
	phi-4	0.12	0.04	—	—	ASR
	gpt-4o-mini-tts	—	—	11.44	0.06	TTS
Spanish	whisper-1	0.02	0.01	—	—	ASR
	phi-4	0.11	0.03	—	—	ASR
	gpt-4o-mini-tts	—	—	10.84	0.07	TTS

Table 6: Evaluation of ASR and TTS models across four languages. ASR performance is reported using Word Error Rate (WER) and Character Error Rate (CER). TTS is evaluated using Mel Cepstral Distortion (MCD) and CER, with the latter obtained via ASR using the whisper-L-v3 model.

Language	R@1	R@5	R@10	P@1	P@5	P@10	F1@1	F1@5	F1@10	MRR
BM25 (T2T)										
Arabic	13.26	35.57	45.12	15.65	8.55	5.45	14.04	13.55	9.62	25.25
Chinese	10.41	26.30	34.95	12.23	6.27	4.15	11.01	9.97	7.34	19.30
English	11.46	27.73	35.13	13.44	6.58	4.16	12.11	10.47	7.37	20.63
Spanish	13.58	33.78	43.37	15.88	8.02	5.17	14.34	12.75	9.15	24.93
Average	12.18	30.84	39.64	14.30	7.36	4.73	12.87	11.68	8.37	22.53
MiniLM-L12-v2 (T2T)										
Arabic	7.13	24.76	36.12	8.82	6.09	4.46	7.68	9.60	7.85	16.87
Chinese	10.26	31.41	44.03	12.34	7.65	5.32	10.94	12.09	9.39	21.93
English	12.99	41.52	56.91	15.67	10.13	6.93	13.87	16.00	12.21	28.08
Spanish	10.70	34.20	48.56	13.07	8.40	5.94	11.48	13.23	10.47	23.64
Average	10.27	32.97	46.41	12.47	8.07	5.66	10.99	12.73	9.98	22.63
text-embedding-3L (T2T)										
Arabic	27.23	65.88	78.73	31.28	15.79	9.55	28.56	25.05	16.84	46.91
Chinese	29.89	70.63	83.11	34.47	16.93	10.06	31.39	26.85	17.75	50.93
English	32.58	75.72	88.03	37.87	18.22	10.72	34.31	28.87	18.89	54.80
Spanish	30.06	71.82	84.57	34.76	17.22	10.27	31.60	27.31	18.10	51.39
Average	29.94	71.01	83.61	34.59	17.04	10.15	31.46	27.02	17.89	51.01
gte-multilingual-base (T2T)										
Arabic	20.65	58.31	74.10	24.50	14.25	9.04	21.91	22.48	15.93	40.09
Chinese	23.37	63.49	79.55	27.48	15.34	9.65	24.71	24.28	17.02	43.80
English	29.91	68.89	82.47	34.51	16.59	9.99	31.41	26.27	17.63	50.38
Spanish	23.50	62.40	79.53	27.88	15.14	9.69	24.93	23.93	17.07	43.98
Average	24.36	63.27	78.91	28.59	15.33	9.59	25.74	24.24	16.91	44.56
NV-Embed-v2 (T2T)										
Arabic	12.80	32.03	41.82	15.06	7.75	5.09	13.54	12.26	8.97	23.91
Chinese	22.75	59.87	74.59	27.24	14.56	9.10	24.22	23.01	16.03	42.47
English	24.65	70.28	87.35	29.45	17.16	10.68	26.22	27.09	18.82	47.48
Spanish	23.60	64.57	81.94	28.19	15.74	10.01	25.10	24.85	17.63	44.96
Average	20.95	56.68	71.42	24.99	13.80	8.72	22.27	21.80	15.36	39.70
CLAP (S2T)										
Ara	0.03	0.20	0.49	0.04	0.06	0.07	0.03	0.09	0.12	0.15
Chn	0.08	0.23	0.35	0.11	0.06	0.05	0.09	0.09	0.08	0.19
Eng	0.10	0.52	0.95	0.11	0.12	0.11	0.10	0.19	0.19	0.34
Esp	0.18	0.42	0.93	0.20	0.11	0.11	0.19	0.17	0.20	0.39
Average	0.10	0.34	0.68	0.11	0.09	0.08	0.10	0.14	0.15	0.27
SpeechT5 (S2T)										
Ara	0.14	0.32	0.49	0.16	0.08	0.06	0.15	0.13	0.11	0.28
Chn	0.06	0.35	0.61	0.07	0.08	0.07	0.07	0.13	0.13	0.23
Eng	0.07	0.16	0.49	0.07	0.04	0.06	0.07	0.06	0.10	0.16
Esp	0.05	0.29	0.68	0.05	0.08	0.08	0.05	0.12	0.15	0.22
Average	0.08	0.28	0.57	0.09	0.07	0.07	0.08	0.11	0.12	0.22

Table 7: Knowledge retrieval performance in two settings: text-to-text retrieval using multilingual text encoders (T2T) and speech-to-text retrieval using multilingual and multimodal encoders (S2T). Retrieval is performed over a fully parallel set of knowledge snippets, enabling direct cross-lingual comparison. The T2T setting uses human-annotated transcripts, while S2T is based on user-recorded audio.

Language	Turn Acc.	Exact Match	OOK Recall
XLM-R <sub>large</sub>			
Arabic	95.39	—	—
Chinese	95.23	—	—
English	96.30	—	—
Spanish	95.93	—	—
Average	95.71	—	—
Llama-3.1-8B-Inst (R@5)			
Arabic	86.75	15.26	5.06
Chinese	92.99	20.73	0.00
English	92.19	18.14	6.33
Spanish	93.30	21.00	13.92
Average	91.31	18.78	6.83
gpt-4.1-nano (R@5)			
Arabic	—	19.96 (20.21*)	3.80 (0.00*)
Chinese	—	22.98 (19.86*)	2.53 (0.00*)
English	—	26.62 (23.02*)	5.06 (14.29*)
Spanish	—	24.87 (21.09*)	5.06 (0.00*)
Average	—	23.61 (21.05*)	4.11 (3.57*)
gpt-4.1-nano (R@10)			
Arabic	—	12.58	3.80
Chinese	—	17.15	0.00
English	—	23.33	1.27
Spanish	—	19.55	3.80
Average	—	18.15	2.22
gpt-4o-nano (R@20)			
Arabic	—	10.85	1.27
Chinese	—	12.28	0.00
English	—	18.72	1.27
Spanish	—	11.03	0.00
Average	—	13.72	0.63
Threshold			
Arabic	—	6.26	48.47
Chinese	—	6.61	46.38
English	—	6.88	63.10
Spanish	—	6.46	43.70
Average	—	6.55	50.41
gpt-4.1 (R@5)			
Arabic	—	34.27*	0.00*
Chinese	—	39.19*	14.29*
English	—	44.29*	42.86*
Spanish	—	39.54*	14.29*
Average	—	39.32*	17.36*
gpt-4.1-mini (R@5)			
Arabic	—	14.94*	0.00*
Chinese	—	23.02*	14.29*
English	—	26.19*	0.00*
Spanish	—	20.56*	0.00*
Average	—	21.18*	3.57*
gpt-4o (R@5)			
Arabic	—	30.05*	14.29*
Chinese	—	35.32*	42.86*
English	—	36.91*	28.57*
Spanish	—	33.04*	42.86*
Average	—	33.83*	32.14*
gpt-4o-mini (R@5)			
Arabic	—	7.91*	0.00*
Chinese	—	11.78*	0.00*
English	—	10.37*	0.00*
Spanish	—	10.02*	0.00*
Average	—	10.02*	0.00*

Table 8: Model performance on retrieval turn classification and knowledge filtering across four languages. Turn classification is evaluated using accuracy (Turn Acc.). Knowledge filtering is measured by Exact Match (EM) and Out-of-Knowledge (OOK) Recall. R@5, R@10, and R@20 indicate the number of top-ranked retrieved snippets considered during filtering. The upper bounds for EM under R@5, R@10, and R@20 are 72.53%, 86.41%, and 92.26%, respectively (averaged across all languages). (\*) Results marked with an asterisk are based on the same randomly sampled 10% subset of the test set, due to the high cost of evaluating the full dataset with these models. This subset contains only 7 OOK turns, leading to high variance in OOK Recall, and should be interpreted with caution.

Language	BLEU	METEOR	ROUGE-L
Llama-3.1-8B-Inst			
Arabic	3.35	15.11	18.24
Chinese	4.86	28.20	0.01
English	5.67	31.45	26.98
Spanish	5.76	26.28	26.22
Average	4.91	25.26	17.86
gpt-4o-nano			
Arabic	3.85	22.93	20.85
Chinese	5.96	33.76	0.00
English	10.99	55.65	36.67
Spanish	6.57	32.77	29.59
Average	6.84	36.28	21.78

Table 9: Model performance on response generation. At time step  $t$ , the model receives as input the dialogue history  $\mathcal{H}_t$ , the ground-truth set of knowledge snippets  $\mathcal{K}_t$ , and a retrieval indicator  $\mathbf{r}_t$ , and generates a system response  $\hat{\mathbf{s}}_t$ .