An end-to-end system for transcription, translation, and summarization to support the co-creation process. A Health CASCADE Study.

Balaskas, Georgios; Papadopoulos, Homer; Loisel, Quentin; Pappa, Dimitra; Efthymoglou, George; Chastin, Sebastien

Published in:
PETRA '23: Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments

DOI:
10.1145/3594806.3596567

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

An end-to-end system for transcription, translation, and summarization to support the co-creation process. A Health CASCADE Study.

Georgios Balaskas
National Centre of Scientific Research "Demokritos" - University of Piraeus
Ag Paraskeyi, Athens, Attiki, Greece
gbalaskas@iit.demokritos.gr

Homer Papadopoulos
National Centre of Scientific Research "Demokritos"
Ag Paraskeyi, Athens, Attiki, Greece
homerpap@dat.demokritos.gr

Quentin Loisel
Glasgow Caledonian University
Cowcaddens Rd, Glasgow, Scotland
United Kingdom
Quentin.Loisel@gcu.ac.uk

Dimitra Pappa
National Centre of Scientific Research "Demokritos"
Ag Paraskeyi, Athens, Attiki, Greece
dimitra@dat.demokritos.gr

George Efthymoglou
University of Piraeus
Pireas, Athens, Attiki, Greece
gefthymo@unipi.gr

Sebastien Chastin
Glasgow Caledonian University - Ghent University
Cowcaddens Rd, Glasgow, Scotland
United Kingdom
Sebastien.Chastin@gcu.ac.uk

ABSTRACT

This paper presents a web service and a deep learning (DL) pipeline that has been developed and implemented as part of the MSCA Health CASCADE project. The purpose of the web service is to provide support, streamline, and enable participatory methods, such as co-creation, in the public health domain. The DL pipeline assists with translation, transcription, speaker diarization, relation extraction, and summarization of audio recordings. This is achieved by removing the need for time-consuming tasks, such as translating and transcribing audio recordings. Additional value is created by extracting implicit relations from the transcribed text and identifying patterns, key themes, and trends. Finally, providing summaries of the transcripts creates a sense of ownership that can improve stakeholder retention in participatory methods.

CCS CONCEPTS

• **Computing methodologies** → **Natural language generation**; **Speech recognition**; **Machine translation**.

KEYWORDS

natural language processing, transcription, speech-to-text translation, speaker diarization, relation extraction, dialogue summarization, co-creation, public health

ACM Reference Format:
Georgios Balaskas, Homer Papadopoulos, Quentin Loisel, Dimitra Pappa, George Efthymoglou, and Sebastien Chastin. 2023. An end-to-end system for transcription, translation, and summarization to support the co-creation process. A Health CASCADE Study.. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '23), July 05–07, 2023, Corfu, Greece.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3594806.3596567

PETRA '23, July 05–07, 2023, Corfu, Greece

ACM ISBN 979-8-4007-0069-9/23/07.
https://doi.org/10.1145/3594806.3596567

1 INTRODUCTION

Public health is crucial for the success of our society, as it entails the promotion of a better standard of living. It encompasses a wide range of activities, from health education and monitoring to promoting healthy, active lifestyles and more [1, 10, 28]. After a global pandemic, its significance has only become more evident during the past few years. We are currently in a pivotal decade, with numerous environmental, economic, and medical challenges that are adversely impacting public health [18–21].

In addressing these challenges, participatory methods have been proven as effective approaches [13]. They involve stakeholders, including individuals, communities, and organizations, to engage in the development and implementation of interventions and policies that can be used to address public health issues [14, 15]. Health CASCADE focuses on one such participatory method, Co-creation. This specific participatory method involves knowledge sharing, collective intelligence processes, and shared decision-making, which enables stakeholders to contribute their perspectives and expertise as equals, thereby increasing the likelihood of successful outcomes [15, 28]. However, co-creation is facing challenges. For example, it is a time-consuming process that prevents its implementation in potentially useful situations [8, 24].

Machine Learning (ML) and, more recently, Deep Learning (DL) have shown enormous potential in successfully addressing issues in a large number of fields. It has enabled people to accurately and rapidly identify cancer from photos and medical images [26, 30]. To accurately and non-intrusively identify energy consumption in households[3, 25]. It has drastically improved our ability to handle most language tasks, from translation to question answering and text generation [9, 16, 23, 27]. We can see the impact of ML and DL

in every aspect of our society.

Deep learning techniques have the potential to enable and improve co-creation by providing tools that can bridge the gap between different stakeholders. They can improve and speed up the co-creation process by reducing the time-consuming tasks the co-creators and facilitators need to conduct [15]. They can improve the feeling of ownership the co-creators get after the co-creation sessions by rapidly providing outputs. This has shown improvements in co-creators' willingness to participate and stay committed to co-creation, thus improving the overall process and the final outputs produced [15].

Our goal with this paper is to support the facilitators of the co-creation process by reducing their workload, enabling the co-creators by improving their sense of ownership, and providing them with a rapidly available output after each co-creation session.

Thus, we propose using Natural Language Processing (NLP) techniques to create a secure and trusted online service that the co-creators and the facilitator can use to upload an audio or a video recording from the co-creation session. The prototype of the service, which can be found at **https://cocreatewithai.eu/** extracts the audio and then outputs the following:

- **An accurate transcription of the recording in its original language**. The transcription result is a `pdf` document. The transcription is broken down into different sentences with the speaker of each sentence identified using the following format `<| Speaker1:, Speaker2: |>`. This format maintains the speakers' anonymity while providing a more precise representation of the conversation.
- **English Translation when the language of the recording is not English**. The user is asked to provide the language of the recording. If the language is not English, the recording is first transcribed in the native language and then translated and transcribed into English. Both are provided to the user in a `pdf` document.
- **A visual representation of the extracted relations.** The entities and the relations in the transcribed English text are extracted as a triplet of relations between entities. They are then visualized and presented in an interactive `html` file. The triplets are also added at the end of the `pdf` document that contains the transcription for posterity.
- **Summaries of the dialogues**. The transcription of the recording is broken down into segments. These segments are then summarized and added to the `pdf` document.

Privacy concerns and data control are also essential considerations due to the nature of the data created during the co-creation process and the public health focus of the co-creation sessions conducted within Health CASCADE. Co-creators discuss personal information that has to stay private and secure at all costs. As such, co-creators and facilitators must be able to retain control over the data to ensure its proper use and prevent it from being misused or mishandled. Appropriate measures must be in place to ensure that the data are only used following the wishes of the co-creators and then deleted.

The rest of the paper is organized in the following manner. The Web Service Architecture and its components are described in Section 2. The deep learning Pipeline Architecture and its different stages, models, and evaluation are described in Section 3. The impact of the service, its limitations, and proposed future improvements are described in Section 4.

2 WEB SERVICE ARCHITECTURE

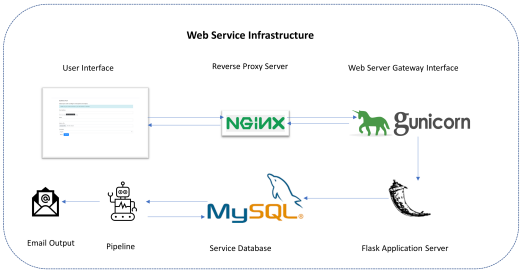

Figure 1: Overview of the web service overall architecture

In this section, we describe our approach to developing the aforementioned web service and pipeline. The web service allows the users to upload their recordings to our secure servers, where our custom DL pipeline processes their data. The aforementioned outputs that the DL pipeline has generated are then emailed to the user.

Our web service aims to improve the flow of the co-creation process, ease the facilitators' burden, and enable the co-creators by providing them with a concrete output of their participation effort. As such, the web service has to be simple to use to improve the experience of the co-creators.

2.1 Web Service Interface

In this section, we present the Interface of the web service. Our service's users are co-creators who can have very diverse backgrounds. To accommodate all possible co-creators, we have designed a very simplistic Interface for the web service to ensure everyone can use it with minimal effort. Additionally, we have created a video tutorial that can be found below the Interface to assist the users through the process.

The service interface consists of a single upload form with the following fields.

- **An email field**
- **A name field**.
- **A file upload field**.
- **A language dropdown**.

To ensure the server's safety and security and our co-creators data. Before we accept an upload request, we check that the files

to be uploaded are valid sound or video files. Additionally, we sanitize the `name` and `emails` the users provide to ensure they are not malicious entries. If these checks pass on the client side, we conduct them again on the server side before saving any files on the server or adding any entries to the database.

2.2 Recordings

The service allows users to upload audio and video file recordings through the website. All widely used video and audio formats are accepted. The service accepts video and audio files of up to 3.5 GBs in size. The audio is extracted from the uploaded recordings, and the format of the files is changed to a suitable `wav` format using the `ffmpeg` Linux library. Then, the files are passed as inputs through the pipeline that generates the aforementioned outputs. Finally, the uploaded files are securely stored on the server and deleted after the pipeline analyses them to ensure the users' privacy and their data.

2.3 MySQL Database

If the input passes the security checks in place and after the recordings have been saved in the appropriate format. The uploaded information is saved in a `MySQL Database` (DB). The DB contains two tables, a table used to implement a queuing system and a table used to log statistical information.

The queueing table saves the `email, name, file path, and language` that the user entered into the upload form. This information is then passed on to the pipeline in a first-in, first-out fashion. After the pipeline successfully produces its outputs, the entry is deleted from the table. Additionally, if more than 24 hours have passed since the entry to the database, the entry is deleted together with the file recording saved in the file path.

The statistics table saves `Id, filepath` fields and five fields that represent the five stages of the pipeline. These fields are then populated with values of 0 to denote that no failures have occurred. If a part of the pipeline fails, the number is incremented by one, and the pipeline processing restarts for up to three attempts. If three fails are saved in the same field, the entry is deleted from the queue table, and the recordings and any generated output are deleted from the server.

2.4 Pipeline

The pipeline handles all the processing of the information in the recordings and the production of outputs. Additionally, it handles sending the email with the results to the users and deleting the recordings, output files, and appropriate database entries. More information on the processing and analyses can be found in the third chapter.

2.5 Web servers

To host our Interface and the service's upload functionality, we use a `Nginx` web server as a reverse proxy. The `Nginx` server handles the requests to and from the client. In addition, the `Nginx` server uses Transport Layer Security(TLS) certificates to improve data security during the communication between the client interface and our servers. Before a request is accepted by the `Nginx` server, it must pass an origin check to ensure that the request originates from the service interface.

The `Nginx` server then sends the request to a Web Server Gateway Interface (WSGI) HTTP server, `Gunicorn`.

The WSGI handles all the interfaces between the `Nginx` server and the `Flask` application. In addition, it handles the sessions and the workers and ensures tasks are restarted in case of failure.

Finally, a Flask application is used to check if the files uploaded are in the correct format and if the user inputs are sanitized and are appropriate to be added to the database. Additionally, it handles the conversion of the files to the correct format and rejects upload requests when the aforementioned criteria are unmet.

3 PIPELINE ARCHITECTURE

In this section, we present that the pipeline employed by the service consists of five stages, in addition to the loading of the sound recordings and the consolidation of the outputs generated, and employs multiple different deep learning models.

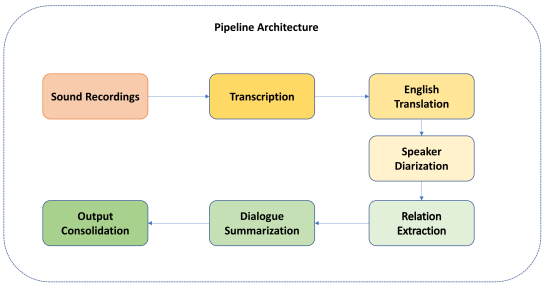

Figure 2: Overview of the pipeline architecture

Initially, the recordings are transcribed in their native language. In addition, if the recording language is not English, the recordings are transcribed and translated using Speech-to-Text Translation to English. [22] The resulting output is a time-stamped text of the transcription and, where applicable, a translated transcription in English. [22]

In the second stage, the audio recording is passed to a model pipeline that uses multi-label classification for the task of speaker diarization. Speaker diarization aims to recognize and categorize distinct speakers in the recording and subsequently includes their corresponding identities to the transcribed and translated text. [4, 5] This process aligns the time stamps of the speakers to those of the transcribed text. The speakers are pseudonymized using the following convention `<| Speaker1:, Speaker2: |>`, etc.[4, 5]

Following this, the transcribed text is subjected to a relation extraction model to identify the entities mentioned in the text and the implicit relations between them. The model used is an auto-regressive model that outputs all triplets present in the input text. [6] This model uses the BART-large model for its base. [16]

The pipeline's last stage involves segmenting the text into smaller segments, which are then passed to a BART model that has been fine-tuned to generate dialogue summaries between different speakers. [7, 16]

3.1 Automatic Speech Recognition

The nature of our real-life audio recordings means that the sound quality is not optimal. Often the recordings contain a number of difficulties, such as multiple speakers with different accents, multiple languages, and speaking from a distance to the microphone. Models trained using only carefully curated data sets and examples can struggle to produce good results in less-than-ideal conditions. [22]

Weighing these considerations, we have decided to use the *Whisper-large* model [22]. It has been trained on `680,000` hours of multilingual and multitask supervised data collected from various web sources. [22] Whisper uses an end-to-end approach that follows an encoder-decoder Transformer architecture [22, 27]. Input audio is split into 30-second chunks, then re-sampled to `16,000 Hz`. They are then converted into a log-Mel spectrogram representation that is then passed through an encoder. A decoder is trained to predict the corresponding text. Several special tokens are passed to the decoder to indicate different tasks or states, such as`<|nospeech|>`, `<|transcription|>`,`<|translation|>` [22].

Whisper provides time-stamp-aware audio transcription capabilities and accurate speech-to-text translation capabilities for 97 languages. Public health problems are prevalent everywhere, and the impressive capabilities of Whisper in transcribing and translating such a large number of languages out of the box make it an optimal choice for our pipeline.

3.2 Speaker Diarization

To enhance the pipeline results, we need to identify the speakers. This allows us to map the opinions of different stakeholders accurately and enables us to create cleaner transcripts and improve the summarization of the dialogues in the recordings.

To achieve accurate time-stamp-aware speaker diarization, we used an end-to-end neural speaker diarization approach (EEND) [4, 5]. Traditionally, speaker diarization was done in three steps: voice-activity detection, speaker change detection, and overlapped speech detection. However, this approach suffered in terms of robustness when met with non-optimal real-life conditions [4, 5]. Combining the different tasks in an EEND pipeline makes it possible to jointly optimize their hyper-parameters and minimize the diarization error rate[4, 5].

In our pipeline, we use the Pyannote pre-trained EEND [4, 5]. It has been trained as a multi-label classification problem using permutation-invariant training[4, 5]. The audio is split into small chunks of 5 seconds with a sampling rate of `16,000` Hz. By processing short audio chunks, the number of different speakers in each chink is smaller and less variable, making the speaker diarization task easier[4, 5].

3.3 Relation Extraction

Co-creation sessions involve stakeholders from different backgrounds, each with specific knowledge and expertise. The information generated during these sessions contains valuable insights, and perspectives [15, 28]. Often, the relation between these insights is not readily apparent.

By extracting relations from the English transcribed text we have generated in the previous steps of the pipeline, we can create a knowledge graph to represent these relations visually. This can enable and improve the co-creation process twofold. (1) It can help the co-creators and facilitators identify key themes, trends, and patterns. Additionally, it can help identify patterns, such as areas of agreement and disagreements between stakeholders, that can help guide future co-creation sessions[6]. (2) It can improve the sense of shared ownership and collaboration between co-creators and, by extension, improve the retention of co-creators within the co-creation process[15].

Co-creation in public health is an enormous field that is hard to map accurately in sets of predefined entities and relations. As a result, it is optimal to use an auto-regressive approach that frames Relation Extraction as a `seq2seq` task, similar to translation, by leveraging a `BART` pre-trained model[6, 16]. When training a `seq2seq` model for a translation task, the encoder receives the text in the original language. Then, the decoder receives the text in the translated language and outputs a prediction[6]. By representing Relation Extraction as a translation task, we can provide the raw text that implicitly contains the entities and their relations and output a set of triplets[6].

To achieve this, we use a `BART` model that has been fine-tuned on the `REBEL` dataset[6]. The model's output is in the form of a triplet `<head - relation - tail>`.

3.4 Dialogue Summarization

The information generated during the co-creation sessions comes from the dialogue between the different co-creators. By providing concise and informative summaries of the conversations between the co-creators, we improve their sense of ownership and streamline the co-creation process. The facilitators can rapidly provide the co-creators with the key information generated during the co-creation sessions. It is also a very useful tool for resuming conversations and topics of discussion from previous co-creation sessions.

To successfully generate dialogue summaries, we identified two relevant datasets `DialogSum`[7], `SAMSum`[12]. We identified a fine-tuned `BART` model [16, 17] that achieves state-of-the-art results in

the aforementioned datasets. Using a pre-trained `BART-large-xsum` [16] as the best starting point. BART models have consistently shown SOTA performance in summarization tasks, and `BART-large-xsum`[11] has already been fine-tuned for abstractive summarization. By further fine-tuning the model on the two identified datasets, its performance is further improved in the task of dialogue summarization[17].

To apply the model, we use the English speaker-aware transcription of the recordings. First, we split the recordings into smaller segments and then used the model to generate summaries for the individual segments. Finally, we combine the segments to create an overall summary of the co-creation session. The datasets,`DialogSum`[7], `SAMSum`[12], the model has been fine-tuned on, contain short dialogue segments as seen in Table 1.

Datasets	Domain	Dialogues	Tokens/dialogue
SAMSum	multiple	16,369	94
DialogSum	multiple	13,460	131

Table 1: Average Length of tokens per dialogue and turn on Dialogue Summarization Datasets [7]

However, co-creation sessions can last up to an hour and, in some cases, more. This means that the generated transcription is often large enough that important information is lost during the summarization process if the text is not appropriately segmented. To address this limitation, we split the transcribed text into segments of $Length = 200$ words. If the overall length of the transcribed text is less than $Length = 200$, the summarization stage of the pipeline will be skipped.

3.5 Outputs

Once the pipeline has finished its analyses, we combine the different outputs into two files. The text output is combined into a *pdf* file with the following format:

- **An accurate speaker-aware transcription of the recording in its original language**
- **An accurate speaker-aware translation of the recording in English if the original language is not English**.
- **A set of segments of the English transcribed text and a summary per segment** .
- **An overall summary of the English transcribed text**.
- **A Table containing the relation triplets identified**.

The relation triplets we identified are visualized using the `Pyvis` python library into an interactive `html` document.

3.6 Evaluation

The web service and pipeline have been implemented and are presently functional, though ongoing development aims to enhance its functionality. To evaluate the web service, we have conducted preliminary internal testing. The preliminary testing has provided us with valuable insights into the functionality of the service. The results of the preliminary testing can be seen in Table 2.

Task	# Success	# Failures	English	Other
Transcription	121	0	76	45
Translation	45	0	-	45
Speaker Diarization	121	0	76	45
Relation Extraction	121	0	76	45
Dialogue Summarization	120	1	76	45

Table 2: Internal test results of the web service

During the preliminary testing, we were able to identify the following two problems and address them. (1) The testers were able to upload the same file multiple times. However, transcribing the recordings is a time-consuming process. By uploading the same file multiple times, the analyses of other files were delayed due to the limited resources available. To address this issue, we implemented a file identification system to ensure that the same file is not accepted multiple times. (2) The summarization process and the relation extraction process can fail if the length of the text is below a certain number of words.

The end-to-end transcription and summarisation system has been tested using the *Common Voice*[2, 22], *VoxPopuli En*[22, 29], *REBEL*[6], *SaMSum*[12] and *DialogSum*[7] datasets.

Additionally, we have designed a lab testing questionnaire with our colleagues to validate the end-to-end transcription and summarisation system. The co-creators and testers of the system are asked to fill out the questionnaire. The testing aims to validate the accuracy and robustness of the DL pipeline in a controlled environment.

4 DISCUSSION AND CONCLUSION

This paper presented our implementation of a DL pipeline and its web service. The service is built to enable, streamline, and improve the co-creation process. The initial testing methodology we designed provides a reliable and repeatable way to evaluate the accuracy and robustness of the service in a lab setting. The methodology has been used to validate the performance of the service and to identify areas for improvement. Our pipeline outputs material that can be used by the co-creators and the facilitators to speed up the co-creation sessions by using the dialogue summaries we generate to recap previous sessions quickly.

Additionally, it removes the need for manual, time-consuming transcription of session recordings. It removes the need for the translation of documents to English, a prevalent need when there are stakeholders with different linguistic backgrounds or when results need to be widely published. It helps build motivation for the co-creation outputs by providing the co-creators with an immediate result for their effort and commitment. Finally, it provides a visual aid that facilitates the advancement of group work by illustrating the relations among the topics deliberated during the discussions.

We have also identified limitations in our pipeline. First, Speaker Identification is not always optimal, especially in cases of overlapping speakers. Second, the transcription time stamps are not always accurate, which affects the speaker-sentence alignment. Third, relation extraction and dialogue summarization is only available for English text. An additional limitation exists in the number of concurent users. The current prototype of the service has limited computing capabilities, as such it is not able to service a large number of concurent user.

We have identified extra functionality that can be useful to the co-creation process and the co-creators. First, more robust Voice-Activity Detection models would help with better-aligning timestamps to transcriptions. We want to include topic modeling and classification to more accurately provide a list of topics, associated keywords, and which co-creators were for or against a specific topic. We want to generate questions from our summaries and use Question Answering systems and literature provided by the facilitators to answer the questions we generate. Finally, the most time-consuming part of our pipeline is the transcription. Addressing this problem would help massively in extending the service to more co-creators.

ACKNOWLEDGMENTS

This study was funded by the European Union's Horizon 2020 Research and Innovation Program under the Marie Skłodowska-Curie grant agreement 956501. The views expressed in this manuscript are the author's views and do not necessarily reflect those of the funders.

REFERENCES

[1] Qingfan An, Marlene Sandlund, Ragnberth Helleday, Danielle Agnello, Lauren McCaffrey, and Karin Wadell. 2022. Co-creation Practice in the Development of Non- pharmacological Interventions for People with Chronic Obstructive Pulmonary Disease: A Health CASCADE Scoping Review Protocol. https://doi.org/10.5281/zenodo.6684694

[2] Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, 4218–4222. https://aclanthology.org/2020.lrec-1.520

[3] Sotiris Athanasoulias, Stavros Sykiotis, Maria Kaselimi, Eftychios Protopapadakis, and Nikolaos Ipiotis. 2022. A First Approach using Graph Neural Networks on Non-Intrusive-Load-Monitoring. 601–607. https://doi.org/10.1145/3529190.3534722

[4] Hervé Bredin and Antoine Laurent. 2021. End-to-end speaker segmentation for overlap-aware resegmentation. In *Proc. Interspeech 2021*.

[5] Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2020. pyannote.audio: neural building blocks for speaker diarization. In *ICASSP 2020, IEEE International Conference on Acoustics, Speech, and Signal Processing*.

[6] Pere-Lluís Huguet Cabot and Roberto Navigli. 2021. REBEL: Relation extraction by end-to-end language generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*. 2370–2381.

[7] Yulong Chen, Yang Liu, Liang Chen, and Yue Zhang. 2021. DialogSum: A real-life scenario dialogue summarization dataset. *arXiv preprint arXiv:2105.06762* (2021).

[8] Jane Clemensen, Mette J Rothmann, Anthony C Smith, Liam J Caffery, and Dorthe B Danbjorg. 2017. Participatory design methods in telemedicine research. *Journal of telemedicine and telecare* 23, 9 (2017), 780–785.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[10] Agnello DM, Loisel QEA, An Q, Balaskas G, Chrifou R, Dall P, de Boer J, Delfmann LR, Giné-Garriga M, Goh K, Longworth GR, Messiha K, McCaffrey L, Smith N, Steiner A, Vogelsang M, and Chastin S. 2022. Establishing a Curated Open-access Database to Consolidate Knowledge About Co-creation: A Health CASCADE Study Combining Systematic Review Methodology and Artificial Intelligence. https://doi.org/10.5281/zenodo.6817196

[11] facebook. 2023. facebook/bart-large-xsum. https://huggingface.co/facebook/bart-large-xsum/tree/main

[12] Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. 2019. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. *arXiv preprint arXiv:1911.12237* (2019).

[13] Kristoffer Halvorsrud, Justyna Kucharska, Katherine Adlington, Katja Rüdell, Eva Brown Hajdukova, James Nazroo, Maria Haarmans, James Rhodes, and Kamaldeep Bhui. 2021. Identifying evidence of effectiveness in the co-creation of research: a systematic review and meta-analysis of the international healthcare literature. *Journal of public health* 43, 1 (2021), 197–208.

[14] Lisan M Hidding, Mai JM Chinapaw, Laura S Belmon, and Teatske M Altenburg. 2020. Co-creating a 24-hour movement behavior tool together with 9–12-year-old children using mixed-methods: MyDailyMoves. *International Journal of Behavioral Nutrition and Physical Activity* 17, 1 (2020), 1–12.

[15] Calum F. Leask, Marlene Sandlund, Dawn A. Skelton, Teatske M. Altenburg, Greet Cardon, Mai J. M. Chinapaw, Ilse De Bourdeaudhuij, Maite Verloigne, Sebastien F. M. Chastin, and Safe Step and Teenage Girls on the Move Research Groups on behalf of the GrandStand. 2019. Framework, principles and recommendations for utilising participatory methodologies in the co-creation and evaluation of public health interventions. *Research Involvement and Engagement* 5, 1 (Jan. 2019), 2. https://doi.org/10.1186/s40900-018-0136-9

[16] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461* (2019).

[17] Karthick Kaliannan Neelamohan. 2023. knkarthick/MEETING SUMMARY. https://huggingface.co/knkarthick/MEETING_SUMMARY

[18] World Health Organization et al. 2009. Summary and policy implications Vision 2030: The resilience of water supply and sanitation in the face of climate change. (2009).

[19] World Health Organization et al. 2016. Global strategy on human resources for health: workforce 2030. (2016).

[20] World Health Organization et al. 2017. Global vector control response 2017-2030. *Global vector control response 2017-2030*. (2017).

[21] World Health Organization et al. 2022. *WHO global strategy for food safety 2022-2030: towards stronger food safety systems and global cooperation*. World Health Organization.

[22] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356* (2022).

[23] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.

[24] Preeti Sushama, Cristian Ghergu, Agnes Meershoek, Luc P de Witte, Onno CP van Schayck, and Anja Krumeich. 2018. Dark clouds in co-creation, and their silver linings: Practical challenges we faced in a participatory project in a resource-constrained community in India, and how we overcame (some of) them. *Global Health Action* 11, 1 (2018), 1421342.

[25] Stavros Sykiotis, Sotirios Athanasoulias, Maria Kaselimi, Anastasios Doulamis, Nikolaos Doulamis, Lina Stankovic, and Vladimir Stankovic. 2023. Performance-aware NILM model optimization for edge deployment. *IEEE Transactions on Green Communications and Networking* (2023), 1–1. https://doi.org/10.1109/TGCN.2023.3244278

[26] Ahsan Bin Tufail, Yong-Kui Ma, Mohammed K. A. Kaabar, Francisco Martínez, A. R. Junejo, Inam Ullah, and Rahim Khan. 2021. Deep Learning in Cancer Diagnosis and Prognosis Prediction: A Minireview on Challenges, Recent Trends, and Future Directions. *Computational and Mathematical Methods in Medicine* 2021 (Oct. 2021), 9025470. https://doi.org/10.1155/2021/9025470 Publisher: Hindawi.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[28] Maïté Verloigne, Teatske Altenburg, Greet Cardon, Mai Chinapaw, Philippa Dall, Benedicte Deforche, Maria Giné-Garriga, Sonia Lippke, Homer Papadopoulos, Dimitra Pappa, Marlene Sandlund, Margrit Schreier, Karin Wadell, and Sebastien Chastin. 2022. Making co-creation a trustworthy methodology for closing the implementation gap between knowledge and action in health promotion: the Health CASCADE project. https://doi.org/10.5281/zenodo.6817196

[29] Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long*

Papers). Association for Computational Linguistics, Online, 993–1003. https://doi.org/10.18653/v1/2021.acl-long.80

[30] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers. 2017. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* IEEE Computer Society, Los Alamitos, CA, USA, 3462–3471. https://doi.org/10.1109/CVPR.2017.369