1. INTRODUCTION

Communication is a fundamental element that determines the effective functioning of people in society. It is the process of exchanging information that can occur between humans or humans and machines. In an era of continuous technological development, people are surrounded by all sorts of electronic devices, including embedded systems such as everyday devices and medical sensors. As time goes by, these devices become more complex and have more functions, making communication with them more complicated. Therefore, it is essential for people to develop their communication skills in order to effectively operate these devices and take full advantage of their capabilities. An example of algorithms that improve communication is the Speech-To-Text (STT) method of recognizing speech and converting it into written text. It has a number of potential applications, including hands-free operation of devices, which is particularly useful for people with mobility or visual limitations, as it enables them to enter text more easily. STT can also be used for natural language processing in applications and systems, allowing users to communicate with computers and other devices in a more intuitive and natural way. STT can also be designed to recognize and transcribe multiple languages, making it a useful tool for companies and organizations operating in multiple countries or regions[1].

Speech synthesis is the process by which an acoustic signal is generated that mimics human speech. The first step in this process is the identification of words and how they are pronounced through phonetic notation. The notation takes into account the pronunciation only to generate a single word and not the entire utterance because speech does not consist of single words but of a sequence of utterances, taking into account the semantic context and emotions. Speech synthesis should, therefore, be based on lexical, syntactic, as well as semantic analysis. Another important element of speech synthesis is the appropriate prosody of the signal, that is, intonation and the length of speech segments. In addition, it is necessary to consider the style that each speaker has ─how fast they speak, whether it is a raised voice or a whisper, etc. All the variables listed above will affect the output signal[2,3,4].

In contrast, Text-to-Speech processing is a mechanism in which text (a string of characters) is converted into an audio signal[5]. Clarity of speech is one of the most important features to consider and must be generated in real-time. The naturalness of the speech signal, on the other hand, reflects to what extent the generated speech (output signal) resembles human speech. In recent years, significant progress can be seen in the area of the issue at hand, which has led to improvements in machine learning models[6]. Unlike the traditional methods for synthesizing human speech, deep neural networks have proven to be most effective in learning linguistic features from training data. However, deep models require extensive data in the training process[7,8,9]. Among examples of datasets employed by deep models LibriSpeech[10] and CommonVoice[11] are often used in training and testing. LibriSpeech[10] is a database of about 1,000 hours of speech recordings in English, which were extracted from audiobooks (LibriVox project). The sampling frequency is 16 kHz. The aforementioned data are preprocessed, and segments with noisy transcriptions are filtered out. This database is publicly available, but due to its large size, it is available in three subsets containing about 100, 365, and 500 hours. CommonVoice[11] is a multilingual database that is provided as part of open-source software and was made available by Mozilla. The database contains about 7,000 hours of verified recordings, in more than 60 languages (including Polish). The database includes demographic metadata such as age, gender, and accent.

The main purpose of this study was to develop deep text-to-speech/speech-to-text (TTS/STT) algorithms designed for the Raspberry Pi 4 embedded device using deep learning models. Another aim was to enhance communication between humans and devices (e.g., assistive devices) capable of performing both TTS and STT functions, especially in interactions with healthcare professionals.

The paper presents the design of an embedded device (Raspberry Pi evaluation board) that includes a proposal for these algorithms. A critical review of the literature on the mechanisms of speech signal generation and processing, as well as the techniques used in speech synthesis and speech-to-text conversion algorithms, is shortly recalled. The design assumptions given in the Study Background Section are described along with the implementation of the programs on the embedded device and the methodology of the tests performed. The tests were designed to examine the degree of correctness of the subjects' word recognition and the degree of intelligibility of the speech generated by the device.

2. STUDY BACKGROUND

The increasing use of embedded systems in many fields has made speech signal processing algorithms a part of everyday life. The ability to use low-cost and easily reprogrammable microprocessors has made it much easier to implement the presented algorithms in applications that were not previously related to audio signals. The use cases for the mentioned algorithms can be divided into two main categories: speech synthesis and speech recognition.

The algorithms were designed for use with the Raspberry Pi 4 board (see Fig. 1)[12]. Also, Raspbian, a free operating system based on the Debian framework that allows the developer board to be used as a computer, but with significantly fewer resources, was employed. It is the core system for the Raspberry Pi, while still being optimized for the board's hardware. The Raspberry Pi operating system is being actively improved in terms of stability and performance with a substantial amount of packages offered by Debian. The featured operating system has a PIXEL (Pi Improved Xwindows Environment, Lightweight) desktop environment, which looks similar to typical desktops such as those in Windows or macOS.

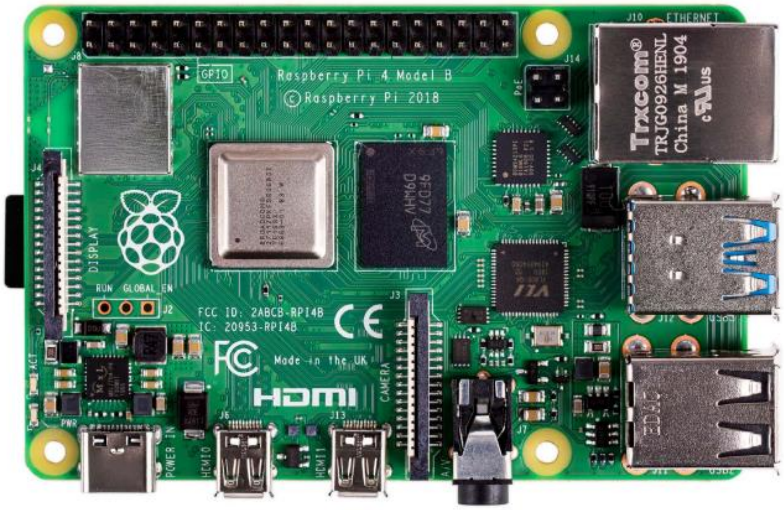

Fig. 1. Raspberry Pi 4B board[12].

The speech-to-text conversion algorithm used in this study is TensorFlowTTS[13], which provides architectures for speech generation models such as Tacotron2[14], Melgan, and Fastspeech[15]. The Tacotron2 architecture was used, which is notable for its performance and the good quality of the generated recordings. This is an encoder-decoder model that generates a Mel spectrogram from the text, enhanced by the WeveGlow[14] model for waveform conversion. To use the model on the target platform, it was first necessary to perform preprocessing of the database, training of the model, and compression of the model in such a way that it is optimized as much as possible for an embedded system such as Raspberry Pi. In Fig. 2, the block diagram of Tacotron 2 is shown[15]. It is important to note that 2D signal representation is used in the form of a Mel spectrogram in the signal processing path. Several convolutional layers, as well as LSTM (Long short-term memory) neural networks, are provided in this architecture, including a bidirectional LSTM[15].

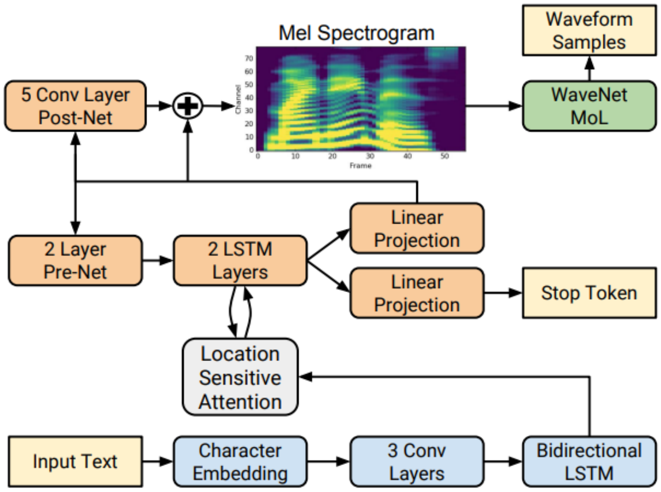

Fig. 2. Tacotron architecture[14].

3. EXPERIMENTAL SETUP

To train the neural network model, it was necessary to select a suitable database, which was later used to train and validate the correct operation of the model. For this purpose, the LJSpeech[16] database compatible with the algorithm was chosen, which became a reference in most speech-to-text conversion systems. The LJSpeech database is a publicly available database released as a part of the LibriVox project[16]. The main parameters of the database are presented further on.

80 synthesized sentences were prepared based on medical and everyday language employing the TTS algorithm. Two types of algorithms were tested: text-to-speech (TTS) and speech-to-text STT). A modified version of the WaveNet vocoder architecture was used to convert spectrograms into samples of the speech signal in the time domain. A mixture of logistic distributions (MoL) of 10 elements was used to generate 16-bit samples at 24 kHz. Finally, the Wavenet output was passed through ReLU activations and a linear projection to predict the parameters for each mixture component[15].

A survey was also prepared for subjective evaluation. In the subjective tests, it was necessary to use headphones, a microphone as well as a sound card of good quality. In the study performed, the microphone was Genius MIC-01C, and the sound card was ugo UKD-1086. Respondents assessed several speech features such as intelligibility, tempo, naturalness, and accentuation.

TTS: TensorFlowTTS was used, which provided speech generation model architectures (e.g., Tacotron2, Melgan, FastSpeech). The model employed was Tacotron2. Other setup details were as follows: compression of the model based on the TensorFlowLite library; database: LJSpeechLibriSpeech[16]: 13,000 audio recordings (in FLAC format) with transcription; sampling frequency: 22,050 Hz.

After preprocessing the data – the database consisted of recordings of audio signals, transcriptions, and Mel-scale spectrograms that were derived from the speech signals.

STT: The publicly available STT engine was used: DeepSpeechSpeech signal recognition in two modes: "offline” and "real-time processing”; model compression based on TensorFlowLite library; database: LJSpeech.

Validation: Tacotron2: a string was introduced as inputDeepSpeech: recordings were in WAV format, and the sampling rate was 16 kHz.

Apart from the everyday language contained in the LJSpeech test database, due to the lack of specialized language, specifically medical terminology, a selection was made of individual excerpts from the ”Gray's Anatomy" audiobook series made available through the LibriVox project. The selected excerpts were characterized by the diversity of male and female voices. Additional editing operations, mainly trimming of the recordings, were performed in the Audacity software.

A. OBJECTIVE EVALUATION

It should be noted that there are several measures used in the automatic speech recognition (ASR) and machine-based transcription area. Their applicability depends on the specific task and goals of the evaluation. The most commonly applied is WER (Word Error Rate), defined as the percentage of words in the output that are incorrect or missing compared to the reference, as it enables the assessment of the overall accuracy. MER (Match Error Rate) refers to the precision in producing the correct words in the output. Similarly, if the evaluation aims at assessing the quality of the output in terms of how well it preserves the information from the reference, then WIP (Word Information Preserved) is to be used. In contrast to that is WIL (Word Information Loss) as it answers to what extent information contained in the reference is lost. In tasks related to ASR, especially at the character-level processing, CER (Character Error Rate) should be applied. There are two additional measures used in ASR and machine-based translation, i.e., RPER (Reference Position-independent word Error Rate) and HPER (Hypothesis Position-independent word Error Rate) that are similar to some extent to WER, but they assess word errors irrespective of their positions.

Since the experiments performed have several goals, i.e., ASR, STT, TTS, etc., that is why the above-mentioned measures were employed to check the efficiency of the algorithms implemented on the Raspberry Pi 4 board. They were as follows: **WER**, **MER**, **WIL**, **WIP**, **CER**, **RPER**, and **HPER**.

Recordings included diverse accents, proper names, and specialized language (medical).

After performing speech processing, several measures were calculated for the outcome of STS and TTS. As already mentioned, several objective quality measures were evaluated, namely **MER, WER**, **WIP**, **CER**, **RPER**, and **HPER**. The evaluation outcomes are contained in Tables 1 and 2.

Table 1. MER values for everyday and specialized (medical) language.

Speaker's gender	**Language category**	**MER [%]**
female	accent	13.51
	names	8.28
	everyday	4.14
	medical	16.57
male	accent	12.73
	names	8.19
	everyday	3.82
	medical	16.86

Table 2. Summary of tested parameters for the STT test set (specialized (medical) language).

Speaker's gender\Measure	**WER [%]**	**MER [%]**	**WIP [%]**	**WIL [%]**	**CER [%]**	**RPER [%]**	**HPER [%]**
female	12.67	12.58	78.53	21.47	6.95	10.03	9.18
male	10.88	10.76	81.66	18.34	6.44	9.29	9.29
summary	11.84	11.74	79.97	20.03	6.71	9.67	9.23

Analyzing the data summarized in Table 1, it can be concluded that for the category of everyday language, both groups received the lowest value of MER. For MER, the lower the score, the more favorable the evaluation of the speech recognition system. Again, the worst-performing category in this comparison is specialized (medical) language. For the categories of proper names and accents, the gender differences are minimal, which may indicate the system's consistent performance in these areas. The category of specialized (medical) language presented the greatest difficulties. This is also seen in Table 2. The overall WER for the female group was

12.67%. For the categories of specialized language and proper names, the male-voiced recordings received a higher WER, although they scored better overall, at 10.88%. For the entire data set, the WER was 11.84%.

Since WIP is defined as the percentage of words that were correctly predicted by the speech recognition system, the higher the score, the better. Examining WIP, it can be concluded that the primary language category is most effectively recognized, regardless of the gender of the speaker. As in the previous measures, the system does well with the category of proper names (WIP > 85%) and accents (WIP ~ 79%). Specialized language remains the most difficult.

As already said, WIL can be defined as a value indicating the percentage of words incorrectly predicted by the speech recognition system. It is assumed that the lower the score, the better. Even though they are opposite in their meaning, WIL is directly related to the WIP measure. Comparing these two measures, it can be seen that the system is consistent in performance. Although the results of WIP can be described as high, information is still lost, including keywords that can change the meaning of an entire sentence. Most keywords are lost for medical terminology, but they were also evident in individual sentences in other categories.

In addition, CER, which determines the percentage of incorrectly predicted characters, was examined. For this parameter, the lower the score, the better, with a score of 0 being a perfect score. Analyzing detailed results, it can be said that for the groups of everyday basic language and proper names, CER achieves less than 4%, while for the other two categories, the results reach more than 9%. This is due to the under-training of the algorithm for specialized language. The given words that do not belong to colloquial speech present worse by about 5 percentage points. The CER for both genders presents similarly, i.e., 6.95% and 6.44% for the female and male groups, respectively. The entire data set thus received 6.71%.

RPER and HPER, are position-independent (PER). The PER compares words in hypotheses with reference sentences, and the score should be less than or equal to the WER. HPER refers to words occurring in hypothesis sentences that do not occur in reference sentences. RPER, on the other hand, refers to words occurring in the reference sentence that do not include the words in the hypothetical sentence. The results for each category are less than the WER values calculated. Among the categories, basic language has the lowest RPER and HPER, while specialized language has the highest. The largest gender differences can be observed in the HPER index for the accent category. Overall, for the female gender, the coefficients studied were: RPER equal to 10.03%, HPER = 9.18%. For the male gender, both parameters were 9.29%. For the overall data set, an RPER of 9.67% and HPER of 9.23% were obtained.

As seen from the above discussion, it is valuable to compare measures as they complement each other and may indicate whether the speech recognition system is consistent in its performance.

B. SURVEY – SUBJECTIVE TESTS

In addition, a set of sentences generated by the system was evaluated subjectively. The survey consisted of six recordings (in order from the shortest to the longest) and a "summary" recording. Users rated intelligibility, tempo (pace), naturalness, and accentuation on a 5-point Likert scale (1-very bad, 5-very good). The statistical parameters considered in this analysis were mode, median, spread, and quartile.

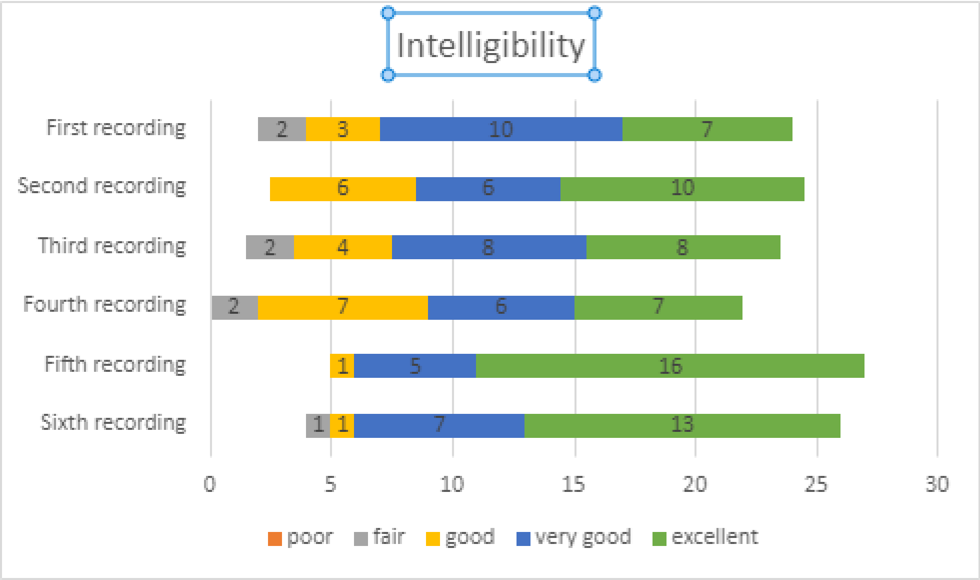

Fig. 3. Results of the survey analysis with regard to intelligibility.

Analysis of the results regarding the intelligibility of the recordings presented to the respondents indicates that most recordings were well understood. The best ratings in this category were given to the longest recordings (fifth and sixth). Moreover, analyzing the detailed results, the fifth and sixth recordings have the highest scores, with a median of 5. In contrast, the third and fourth recordings received two modes, which may indicate that respondents' reception of these audio files varies. For all recordings, the quartile interval is 1 or 2, indicating fairly uniform ratings for the measure under study.

In addition, some other speech-related parameters were evaluated, i.e., the speed of speech, naturalness, and accentuation. In the context of generating speech, proper adjustment of tempo is key to achieving the effect of "naturalness”. The best ratings in this category were given to the second, fifth, and sixth recordings. All analyzed recordings scored fairly uniformly in terms of tempo, with slight differences. Examining the median, for most recordings, with the exception of recording five, it takes the value of 4, which indicates good intelligibility. For modes, the dominant value is very good, with the exception of recording no. two and recording no. four–most frequently rated as good. The largest gap in the evaluation was in the case of recording one, indicating a wide variation in ratings for this recording.

In the naturalness category, the best results in this category were achieved by the second and fifth recordings. For most recordings, the median and mode values oscillate around values of 3 and 4, with the exception of the sixth recording (mode = 5), which may indicate that the recordings are perceived as moderately natural.

Accentuation, like tempo, has a key impact on the perception of audio recordings. Correct accentuation improves intelligibility but also aims to emphasize important elements of speech. The best score in this category was received by the fifth recording, immediately followed by the second recording. However, it is worth noting that the recordings were listened to by people whose language used in the recordings was not their native language, so they were more sensitive to this parameter.

4. CONCLUSION

Synthesizing speech in a way that can be understood by electronic devices or other systems can bring many benefits to applications used every day. Different algorithms may have different sets of capabilities, so choosing the right algorithm depends on the specific application. In simple terms, these types of algorithms will act as playback–for example, sensors that continuously measure the temperature of a machine, which notify the user of the measured value by voice every set time interval, or fire protection systems, etc. In other applications, speech generation algorithms are capable of formulating entire sentences or groups of sentences.

Overall, in this study, the rating of the recordings evaluated subjectively oscillated around a score of good (4) in all assessed categories. Further, difficulty appeared in determining the "worst" and "best" recordings. No direct correlation was noted between the length of the recording and its quality. However, both the longest and shortest recordings received high marks (5) in each category.

Moreover, the subset containing male recordings is better recognized than recordings containing female voices. The Word Error Rate (WER) measure for the entire test data is 11.84% (state-of-the-art (SOTA) refers to 7.06%; other similar studies report WER of 14%). Other measures from recent studies (the year 2020) on similar Speech-to-Text (STT) systems (IBM, Google, Wit) show results twice as high as those obtained in this study. However, the most important conclusion derived from this study is that Text-to-Speech (TTS) and Speech-to-Text (STT) based algorithms need to be trained in the context of applications using specialized (medical) language.

Since medical language recognition can enhance the efficiency and accuracy of clinical documentation by automatically transcribing spoken patient-doctor interactions or converting handwritten notes into digital text, thus this area, i.e., medical natural language processing (MNLP), needs to be thoroughly investigated. This is demonstrated by the recent paper by Boonstra et al.[18], showing that MNLP may revolutionize healthcare, allowing for broader applicability and accessibility of information. It should, however, be pointed out that this review paper regards large language models that are required to be created.

ACKNOWLEDGMENTS

This work is supported by the Polish National Center for Research and Development (NCBR) project: “ADMEDVOICE-Adaptive intelligent speech processing system of medical personnel with the structuring of test results and support of therapeutic process,” no. INFOSTRATEG4/0003/2022.

REFERENCES

[1] D. Yu and L. Deng, “Automatic Speech Recognition: A Deep Learning Approach”. Springer Publishing Company, Inc. (2014).

[2] Z. Yin, “An Overview of Speech Synthesis Technology,” Eighth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC), 522–526 (2018). doi: 10.1109/IMCCC.2018.00116.

[3] X. Wu, “Research on the Introduction to Models used in Speech Recognition,” Proceedings of the 8th International Conference on Humanities and Social Science Research (ICHSSR 2022) 2014–2020 (2022). doi: https://doi.org/10.2991/assehr.k.220504.363.

[4] H. Zen, A. Senior and M. Schuster, “Statistical parametric speech synthesis using deep neural networks,” IEEE International Conference on Acoustics, Speech and Signal Processing, Vancouver, BC, Canada, 7962-7966 (2013). doi: 10.1109/ICASSP.2013.6639215.

[5] S. Dobrisek, J. Gros, F. Mihelic, and N. Pavesic, “HOMER: a voice-driven text-to-speech system for the blind,” ISIE ’99. Proceedings of the IEEE International Symposium on Industrial Electronics (Cat. No.99TH8465), **1**, 205–208 (1999). doi: 10.1109/ISIE.1999.801785.

[6] J.-C. Junqua and J.-P. Haton, “Fundamentals of Automatic Speech Recognition, in Robustness in Automatic Speech Recognition: Fundamentals and Applications,” Boston, MA: Springer US, 73–124 (1996). doi: 10.1007/978-1-4613-1297-0_3.

[7] A. Agarwal, T. Zesch, “German End-to-end Speech Recognition based on DeepSpeech,” KONVENS conference, October (2019), Friedrich-Alexander-Universität Erlangen-Nürnberg, Germany.

[8] P. Bell, J. Fainberg, O. Klejch, J. Li, S. Renals and P. Swietojanski, “Adaptation Algorithms for Neural Network-Based Speech Recognition: An Overview,” IEEE Open Journal of Signal Processing, **2**, 33-66 (2021). Doi: 10.1109/OJSP.2020.3045349.

[9] O. Nazir and A. Malik, “Deep Learning End to End Speech Synthesis: A Review,” 2021 2nd International Conference on Secure Cyber Computing and Communications (ICSCCC), Jalandhar, India, 66-71 (2021). Doi:10.1109/ICSCCC51823.2021.9478125.

[10] LibriSpeech dataset, www.openslr.org/12 (accessed: December ‘2023)

[11] Common Voice dataset, www.commonvoice.mozilla.org/en/datasets (accessed: December ‘2023).

[12] Raspberry Pi https://www.raspberrypi.com/products/raspberry-pi-4-model-b/specifications/ (accessed: December ‘2023)

[13] TensorFlowSpeech, www.github.com/TensorSpeech/TensorFlowTTS (accessed: December ‘2023).

[14] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z.Yang, Z. Chen, Y. Zhang, Y. Wang, RJ Skerry-Ryan, Rif A. Saurous, Y. Agiomyrgiannakis, Y. Wu, “Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions” (2018). https://doi.org/10.48550/arXiv.1712.05884.

[15] Y. Ren, C. Hu, X. Tan, T. Qin, S. Zhao, Z. Zhao, T.-Y. Liu, “FastSpeech 2: Fast and High-Quality End-to-End Text to Speech” (2022). https://doi.org/10.48550/arXiv.2006.04558.

[16] LJ Speech Dataset, LibriVox project, www.keithito.com/LJ-Speech-Dataset/ (accessed: December ‘2023).

[17] A. Morris, V. Maier, P. Green, Phil,” From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition”, 10.21437/Interspeech.2004-668 (2004).

[18] M. J. Boonstra, D. Weissenbacher, J. H. Moore, G. Gonzalez-Hernandez, F. W. Asselbergs, “Artificial intelligence: revolutionizing cardiology with large language models,” European Heart Journal, **45**, 5, 332–345 (2024). https://doi.org/10.1093/eurheartj/ehad838