Journal of Information and Data Management, 2026, 17:1, doi: 10.5753/jidm.2026.5737

Enhancing Automatic Speech Recognition Medical Transcriptions

Yanna Torres Gonçalves [**Universidade Federal do Ceará** | ***yannatorres@alu.ufc.br***]
João Victor B. Alves [**Universidade Federal do Ceará** | ***joaovba2002@alu.ufc.br***]
Breno Alef Dourado Sá [**Universidade Federal do Ceará** | ***brenoalef@alu.ufc.br***]
José A. Fernandes de Macedo [**Universidade Federal do Ceará** | ***jose.macedo@insightlab.ufc.br***]
Ticiana L. Coelho da Silva [**Universidade Federal do Ceará** | ***ticianalc@insightlab.ufc.br***]

Campus do Pici, Bloco 952, Centro de Ciências, Departamento de Computação - Pici, Fortaleza - CE, 60440-900.

Received: 24 March 2025 • **Published:** 13 March 2026

Abstract Automated Speech Recognition (ASR) systems can reduce cognitive load and improve efficiency in medical documentation. This study evaluates Whisper and Wav2Vec2 PT for transcribing medical histories in Brazilian Portuguese. Using real audio-text pairs recorded by specialists and nonspecialists, we assess model performance across speaker profiles. We explore decoding with n-gram language models and post-processing with a BERT-based classifier to correct common spelling errors. Additionally, we apply large language models (LLMs) for text style transfer (TST), converting transcriptions into structured medical anamneses through prompt-based methods. Results show that Whisper outperforms Wav2Vec2 PT overall. The BERT-based correction model improves transcription accuracy, especially when applied after normalization. Among the LLMs tested, Mistral produced the most consistent and structured outputs. These findings demonstrate the potential of combining ASR with language model enhancements for medical documentation, while also highlighting ongoing challenges in clinical ASR.

Keywords: Medical History, Automatic Speech Recognition, Language Model, Text Style Transfer

1 Introduction

The conventional approach to medical history collection is a time-consuming task that places a considerable burden on healthcare professionals [Chiu *et al*., 2017]. According to Hapvida NotreDame Intermédica[1], documenting a patient's medical history can take up to half of the total consultation time. Despite the effort invested, these records may still lack essential details, increasing the risk of incomplete or inaccurate diagnoses.

Additionally, medical histories often exhibit variations in format, structure, and content, making it challenging to standardize and share information among different healthcare providers. Another difficulty is that, during consultations, professionals may find it hard to retain all critical details shared by patients. These issues emphasize the need for solutions that enhance efficiency, standardization, and accessibility in the medical history documentation process.

To mitigate these challenges, automatic speech recognition (ASR) stands out as a promising alternative. ASR technology enables the conversion of spoken language into written text through computational models, leveraging techniques such as pattern recognition and artificial intelligence [Reddy, 1976]. While pre-trained audio encoders like Wav2Vec2 [Baevski *et al*., 2020; Schneider *et al*., 2019] and Jasper [Li *et al*., 2019] have demonstrated their ability to capture high-quality speech representations, their general-purpose training often necessitates fine-tuning for specialized applications, such as medical history transcription. However, this fine-tuning process can be intricate, requiring domain expertise and a sufficiently large dataset containing audio-text pairs relevant to the specific language and field. In the case of medical histories in Brazilian Portuguese, the scarcity of such datasets makes this adaptation particularly challenging.

Ideally, ASR models should generalize effectively across multiple domains without requiring supervised fine-tuning for each deployment scenario. To investigate this capability in the medical context, we establish a benchmark focusing on real audio and text data from medical histories. We use two datasets, one recorded by nonspecialists and another by specialist medical science students, with a more precise pronunciation. This benchmark enables us to evaluate ASR models in different scenarios, including advanced language models like Whisper.

Our study identified several challenges in Portuguese medical transcription, particularly issues arising from phonetic similarities (e.g., "SS" vs. "S") and silent letters such as "H," which contribute to transcription errors in ASR models. The presence of medical abbreviations (e.g., "FC" for *frequência cardíaca*—heart rate) and measurement units (e.g., "bpm" for *batimentos por minuto*—beats per minute) further complicates the task. Additionally, while a doctor might verbally state, "the patient presents a heart rate of 90 beats per minute," it is common for medical records to document this information in a more concise format, such as "HR: 90 bpm".

These challenges highlight the importance of accurate decoding by ASR models or post-processing of transcriptions for Brazilian Portuguese in the clinical field, where precise documentation of medical histories is essential for legal adherence. To tackle issues stemming from complex medical terminology and improve the transcription of patient history recordings, we perform a comparative analysis of Whisper

[1] `https://www.hapvida.com.br/site/`

and Wav2Vec2. Additionally, to ensure the transcription of clinical terms is both accurate and contextually meaningful, we improve the decoding process of the ASR models by incorporating language models to refine and correct the transcriptions. We also explore post-processing strategies using language models to enhance the accuracy and refinement of ASR transcriptions. There is a difference between decoding and post-processing strategies. The decoding strategy is incorporated into the ASR model, while the post-processing strategy is applied after the model's execution.

Another challenge mentioned before in using ASR models is the lack of consistency in the transcriptions. To the best of the authors' knowledge, there is no universal standard for medical reports. However, doctors generally prefer transcriptions that detail a patient's disease history, medical examinations, and related information. Therefore, in addition to exploring strategies to improve the decoding/post-processing phase, we also investigate how to transform the transcription style produced by ASR models into a format similar to structured anamneses. This is a well-known challenge in Natural Language Processing (NLP) known as Text Style Transfer (TST), which seeks to convert text from one style to another while preserving the original content and ensuring the fluency of the transformed text [Lai *et al*., 2024].

This work significantly extends our previous study [Gonçalves *et al*., 2024] with the following contributions: (1) In addition to exploring decoding techniques using language models, such as n-gram, we investigate an alternative post-processing approach using a BERT-based model for token classification. This model specifically targets common ASR transcription errors involving digraphs, like "SS," "RR," "SC," and silent "H" at the beginning of words. Consequently, the model is trained to predict these spelling errors classes: "SS," "RR," "SC," and "H"; (2) We evaluated the use of LLMs through prompt engineering to convert ASR-generated transcriptions into well-structured anamneses, encompassing comprehensive details of patient conditions, diagnostic examinations, and treatment plans; (3) We expanded the related work section; (4) And finally, we conducted more comprehensive experiments to validate our approach.

The rest of this article is organized as follows. Section 2 presents the main related works. Section 3 explains the data, evaluation metrics, and methods used. Section 4 discusses our experimental results. Finally, Section 5 summarizes this work and proposes future directions.

2 Related Work

We organize this section into two main topics: ASR models and methods for decoding and post-processing (after ASR models) enhancements, and approaches using LLMs for text style transfer.

ASR models and methods for decoding/post-processing enhancements. This section presents an overview of key studies that align with our research. Various ASR techniques have been examined within the medical field, such as the study by [Lee *et al*., 2023], which evaluated the effectiveness of a machine learning-based speech recognition system in reducing the documentation workload for nurses in a psychiatric ward. Conducted at Cheng Hsin General Hospital in Taiwan, the study compared documentation time and error rates between speech recognition and keyboard-based input. The results demonstrated that the system processed 30,112 words in 32,456 seconds, with recognition accuracy improving from 87.06% to 95.07% over four sessions. Despite these advancements, errors persisted, highlighting the need for further refinement to enhance the system's reliability in clinical environments.

Similarly, [Paats *et al*., 2018] investigated the impact of different language models on an Estonian ASR system used in clinical settings. Initially, the system relied on a Gaussian Mixture Model (GMM) for acoustic modeling but later transitioned to a Deep Neural Network (DNN) approach. The adaptation process involved training the acoustic model with domain-specific data and fine-tuning the language model using spoken data. Testing with 11 radiologists dictating 219 reports showed notable improvements, reducing the average word error rate (WER) from 18.4% to 5.8%. While these findings demonstrate a significant enhancement in ASR performance, they also emphasize the continued need for domain-specific adaptation and error reduction.

Enhancing ASR transcription accuracy remains a critical concern due to the risks associated with errors in clinical speech recognition, as noted in studies such as [Chiu *et al*., 2017; Kar *et al*., 2021; Sullivan *et al*., 2022]. Beyond fine-tuning, researchers have explored additional techniques, including integrating language models to refine ASR outputs—a strategy that aligns with our approach in this work.

For instance, [Chiu *et al*., 2017] examined two methodologies for developing speech recognition models: one utilizing recurrent neural networks with connectionist temporal classification (CTC) and another based on Listen, Attend, and Spell (LAS) models. The CTC-based system employed context-dependent phoneme outputs, n-gram language models, and pronunciation dictionaries, with decoding performed via a finite-state transducer (FST). Both unidirectional and bidirectional CTC models were trained, yielding a WER of 20.1%, while the LAS model achieved a slightly lower WER of 18.3%.

Further, [Kar *et al*., 2021] proposed an approach for extracting medical information from audio recordings in critical care scenarios. Their method incorporated multi-style training and noise reduction techniques while integrating medical terminology into the ASR system to improve recognition accuracy. Using MetaMap—an NLP tool designed to map text to clinical concepts within UMLS ontologies—the authors enhanced medical term recognition, ultimately reducing WER by up to 52.27% compared to the baseline model.

[Baevski *et al*., 2020] proposes wav2vec 2.0, also known as Wav2Vec2, adopting a self-supervised learning strategy, simultaneously acquiring discrete speech units and contextualized representations. Its architecture includes a feature encoder that processes raw waveform signals through layers of temporal convolution, layer normalization, and GELU activation. The resulting embeddings are then fed into a Transformer-based context network to capture dependencies over time. The authors' results using a one-hour labeled subset of LibriSpeech for training show the feasibility of ASR

models using limited amounts of labeled data.

Another relevant study, [Sullivan *et al*., 2022], focused on optimizing ASR output by incorporating a 4-gram language model into Wav2Vec2's decoding process. This approach helped mitigate spelling mistakes and unrealistic word sequences, increasing the likelihood of correctly predicting domain-specific terms. An advantage of this method is that it enables probability calculations based solely on text corpora, eliminating the need for additional audio data, as required in fine-tuning strategies.

Whisper [Radford *et al*., 2023] is designed to enhance speech processing by addressing the limitations of unsupervised pre-trained models such as Wav2Vec2. While previous models effectively extract speech representations, they often require fine-tuning for specific ASR tasks due to the absence of a robust decoding mechanism. Whisper employs an encoder-decoder Transformer architecture [Vaswani *et al*., 2017], utilizing sequence-to-sequence modeling to directly generate transcriptions. Beyond ASR, it is also capable of language identification and English translation. With training spanning 680,000 hours of multilingual speech data across 96 languages, Whisper demonstrates strong generalization, reducing the need for dataset-specific fine-tuning and making it a versatile solution for multilingual and multitask applications. However, a notable limitation of Whisper is its fine-tuning process, which only supports audio inputs of up to 30 seconds in duration, restricting its applicability for longer audio recordings.

[Sunkara *et al*., 2020] presents a postprocessing framework for joint prediction of punctuation and truecasing in the medical domain, aiming to allow transcription of clinical dictations and doctor-patient conversations without the need to explicitly say commands like "period" or "add a comma." Their proposed model uses pre-trained masked models, such as BERT, along with linear layers to output two predictions, regarding casing and punctuation. The authors reported an approximated 5% absolute improvement for ground truth texts and an approximated 10% improvement for ASR outputs over their BLSTM baseline model under the F1 metric.

The present study builds on the strategies outlined in previous research. Initially, inspired by the work of [Paats *et al*., 2018], the aim was to adapt the model to the medical domain and the Brazilian Portuguese language. However, fine-tuning the ASR models proved to be unfeasible due to three primary constraints: audio-length limitations, infrastructure limitations, and segmentation limitations. These challenges, which will be further explored later in the study, ultimately prevented the successful fine-tuning of the models.

Similarly, in line with [Kar *et al*., 2021], while creating an ontology of medical terms may seem appealing, it presents significant challenges, such as development complexity and the difficulty of finding Portuguese-language databases for the clinical domain. Therefore, this approach was also not pursued.

Therefore, we experiment with language models to enhance decoding, as suggested by [Chiu *et al*., 2017] and [Sullivan *et al*., 2022]. Furthermore, we compare different language models to assess which are most effective for Portuguese. This approach was chosen due to the wide availability of these pre-trained models and their flexibility. Lastly, we use a similar approach to [Sunkara *et al*., 2020] to detect and correct errors in the transcription of words containing silent "H" and digraphs such as "RR" and "SC".

LLMs for Text Style Transfer. LLM prompting has emerged as a highly promising approach in TST applications [Liu *et al*., 2024]. Despite the ongoing debate about the effectiveness of LLMs, their application often demands a significant amount of labeled training examples, particularly in multilingual and stylistically diverse contexts. Nonetheless, there remains substantial potential to further explore and expand the use of LLMs in TST.

[Mukherjee *et al*., 2024] highlights that while some open LLMs demonstrate promising results, they do not outperform the previous state-of-the-art methods. Although the performance of open LLMs on prompting is relatively weaker, fine-tuning leads to significant improvements. The authors experimented with zero-shot prompts, few-shot prompts, and fine-tuning but focused solely on base models for fine-tuning, excluding chat-based and instruction-tuned models. Their experiments were conducted on TST tasks for sentiment transfer and text detoxification.

[Lai *et al*., 2024] proposes a novel framework to guide LLMs in performing TST by leveraging style-specific neurons. The process begins by inputting both source- and target-style texts into the LLM to identify neurons uniquely activated by each style, based on their activation values. Neurons activated by both styles are labeled as overlapping neurons, and their exclusion during style-specific neuron selection is essential, as their presence may hinder the generation of text in the target style. [Lai *et al*., 2024] demonstrates that deactivating neurons exclusive to the source style—while excluding overlapping neurons—enhances style transfer accuracy, albeit with some impact on sentence fluency.

In this paper, we follow a different line of investigation by focusing on instruction-tuned LLMs and domain-specific style transfer, rather than base models or neuron-level interventions. Specifically, we evaluate the efficiency of different LLMs in the task of formatting medical anamneses, a highly structured and sensitive form of text. Our work expands on the prompting strategies explored by Mukherjee *et al*. [2024], but diverges by applying them to instruction-based models and to a specialized domain rather than general TST tasks. While we do not explore internal activations as in Lai *et al*. [2024], we contribute to the understanding of how prompt engineering and model choice affect reliability and consistency in a critical application.

3 Data and Methods

To address our research questions, we followed a multi-step methodology, summarized in Figure 1. The process begins with data preprocessing, which consists of two key stages: audio standardization and text processing. Next, we use two ASR models, Wav2Vec2 and Whisper, to transcribe the audio inputs into text. To improve transcription accuracy, we introduce an additional step that enhances ASR model decoding by incorporating language models, aiming to reduce syntactic

errors, given the crucial role of precise medical histories. Additionally, a text style transfer step is performed to format the transcription according to the expected patient history structure. Finally, we assess the performance of the ASR models using various evaluation metrics, including WER, Cosine Similarity, and BLEU. We also evaluate the effectiveness of the TST step using a Likert scale, based on assessments by three domain specialists. From these evaluations, we extract descriptive metrics and calculate inter-rater agreement using correlation measures such as Kendall's Tau and Fleiss' Kappa.

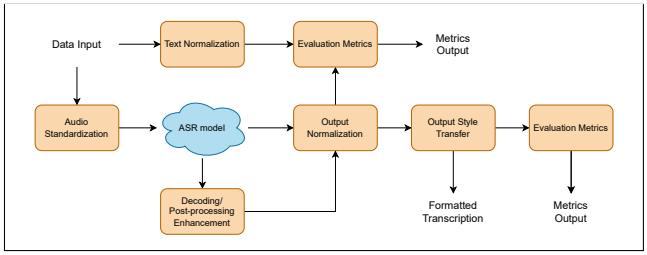

Figure 1. Methodology Overview

3.1 Data source and preprocessing

Due to the lack of a dedicated Brazilian Portuguese medical corpus, we created two datasets of audio-text pairs based on real patient anamneses sourced from Hapvida NotreDame Intermédica's private database. The first dataset, recorded by nonspecialists (two males and one female), consists of 224 pairs representing general patients and was used for ASR model evaluation. The second dataset, recorded by specialists medical science students (two females and one male), contains 382 pairs with longer, more detailed anamneses, precise pronunciation, and validated technical terms. This dataset supports both evaluation and future work, such as fine-tuning.

These datasets allow for performance comparisons between nonspecialist and specialist speakers, as well as between shorter and longer audio recordings. While the specialist recordings show potential for fine-tuning models to better handle Brazilian Portuguese in the medical domain, their extended duration presented challenges. Models such as Whisper are limited to processing 30-second audio segments, and longer audios require significantly more computational resources during training. Therefore, excessively long recordings can hinder both inference and training efficiency.

To ensure consistency and compatibility with ASR models, we applied a text normalization process to all transcriptions and original texts. This process included:

- Converting all text to lowercase and removing leading and trailing whitespace;
- Converting the written form of punctuation to their respective character. E.g., "vírgula" to ",";
- Converting "ºC" to its written form, i.e., "graus celsius";
- Adding spaces between numbers and letters. E.g, "H1N1" becomes "H 1 N 1";
- Converting divisions into their written form. E.g., "2/3" to "2 por 3";
- Converting number separators, eliminating digit grouping, and using "." to separate decimals. E.g., "1.234,56" to "1234.56";
- Converting numbers into their written form. E.g., "42.3" to "quarenta e dois ponto 3";
- Eliminating characters not present in the Wav2Vec2 vocabulary. E.g., "!".

These steps were essential to maintaining uniformity across the dataset and aligning the text with the vocabulary constraints of the ASR model.

3.2 Decoding/Post-processing Enhancement

We observe that ASR decoding can introduce errors, making corrections essential for ensuring text accuracy. To address this, we explore two approaches, both leveraging language models, which capture the fundamental structure and dependencies of language to enhance comprehension and generation. Our investigation focuses on two distinct strategies: the first integrates ASR with an n-gram model, representing a decoding approach. The second employs a BERT-based model specifically designed to correct common ASR transcription errors, such as those involving digraphs and silent "H" from the Portuguese language.

Decoding. The n-gram models calculate the likelihood of word or character occurrences based on preceding context, utilizing extensive textual datasets for training. With nearly 5,000 medical records from Hapvida NotreDame Intermédica, we utilize these texts in our experiments to train the n-gram models. While integrating Wav2Vec2 + n-gram has been previously investigated by [Sullivan *et al.*, 2022], it hasn't been explored in the medical domain. In our experiments, we apply this strategy by leveraging KenLM[2], departing from decoding audio without a language model and enabling the processor to directly receive the model's output logits. This approach, rooted in the decoding process with a language model, enables the processor to consider the probabilities of potential output characters at each time step, thus rectifying any character errors made by the ASR model.

Post-processing. After the transcription, we apply a post-processing step to detect and correct spelling errors encompassing silent "H" and digraphs "SS", "SC" and "RR" from the Portuguese language. The model used for the detection consists of a pre-trained masked language model and a linear layer for token classification. We use BioBERTpt [Rubel Schneider *et al.*, 2020] as the masked language model and align IOB labels accordingly after tokenization, respecting the maximum length of 512 tokens. Following [Sunkara *et al.*, 2020], we only finetune the first 6 layers of BioBERTpt[3] to reduce complexity.

We train the model on a Wikipedia dataset, with 1.28M general texts in Portuguese for training and 160K for validation. Texts have a maximum of 300 words, each automatically processed to, when possible, introduce one of the spelling errors and annotate its presence. We remove the silent "h" from the beginning of the words, change "rr" to 'r', "ss" to "s", and

[2] `https://github.com/kpu/kenlm`
[3] https://huggingface.co/pucpr/biobertpt-all

remove the “s” from “sc” if it is immediately followed by “e” or “i”.

We train the model for 2 epochs, choosing the best result regarding cross-entropy loss on the validation set. Once the model is trained, we use it to submit transcriptions, and correct detected errors following the Portuguese-Language Orthographic Agreement of 1990. For each class, the rules to correct the detected error are as follows:

- H: if the word begins with “a”, “e”, “i”, or “u”, the letter “h” is added at the beginning. E.g., “oje” becomes “hoje”;
- SS: every “s” immediately preceded and succeeded by “a”, “e”, “i”, or “u” is duplicated. E.g., “resecado” becomes “ressecado”;
- SC: if there is a letter “c”, immediately succeeded by the letters “e” or “i”, and not already preceded by an “s”, an “s” is added before the “c”. E.g., “crecimento” becomes “crescimento”;
- RR: every “r” immediately preceded and succeeded by “a”, “e”, “i”, or “u” is duplicated. E.g., “derame” becomes “derrame”.

The spelling correction step can be performed either before (Scenario A) or after (Scenario B) the normalization of the transcription. Figure 2 shows the placement of the spelling correction within the pipeline in both cases. In the example, the transcription of the word “ultrassonografia” is corrected by duplicating the “s” and, as part of the normalization step, the punctuation “:” is removed.

We evaluate this strategy in two aspects. First, we analyze the model’s ability to detect spelling errors by checking the performance on a test set containing 160K texts extracted from Wikipedia. Like the training and validation sets, the test set is automatically processed to introduce errors in the text. We consider the standard token classification metrics: precision, recall, and F_1-score.

After that, we analyze whether using the model to correct transcriptions from the medical domain improves their reliability. We consider both actual and simulated transcriptions of anamnesis. These artificial transcriptions are made by editing the ground truth text to introduce errors, simulating an ASR model that only fails by committing the spelling errors we aim to correct with our model. This ensures that any difference in performance is due to orthography and not other errors ASR models may commit, such as mistaking homophones. We consider both pipeline configurations depicted in Figure 2.

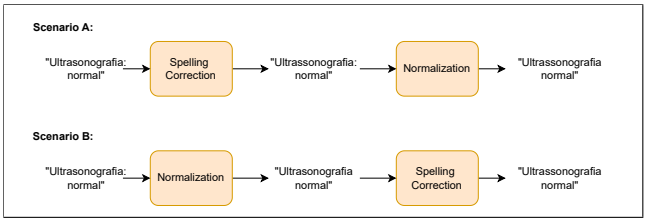

Figure 2. Spelling correction step placement

3.3 Text Style Transfer

Aiming to evaluate the efficiency of different LLMs in the task of formatting medical anamneses, a benchmark was conducted initially. This step allows for a direct comparison of each model’s performance, identifying the one capable of producing the clearest and most consistent formatting while avoiding hallucination given the need for reliability and the sensitive nature of the problem’s scope.

To ensure standardization in the evaluation of the models, a specific prompt was designed, instructing the model to format transcriptions according to a predefined structure. This prompt was applied to three different models: Llama 3.2, Mistral Nemo, and Phi3, chosen for their characteristics and natural language processing capabilities. Each model received the same instructions and was evaluated on formatting 70 real transcriptions, ensuring a consistent comparison. An example of the translated standardized prompt used in this evaluation is presented next, where `"ANAMNESE_DEF"` corresponds to the definition and format of a patient record, and `"transcript"` represents the transcription to be formatted.

```
### Prompt:
<|system|>
You are a formatter for medical records and
anamneses in Portuguese. <|end|>
<|user|>
{ANAMNESE_DEF} Format the following text in
the anamnesis format. If there is not enough
information for a topic, insert ``No information''
or ``Not provided'': <{transcript}>
<|end|>
<|endoftext|>
### End
```

The formatted anamneses were then evaluated by three healthcare students, who rated them from 1 (poor) to 5 (excellent). They could also provide comments on errors, such as missing or misplaced information. To analyze the results, we computed metrics such as mean, minimum, maximum, and standard deviation of the ratings. Kendall’s Tau and Fleiss’ Kappa were applied to assess the consistency and agreement among evaluators, ensuring the reliability of the evaluation.

3.4 Evaluation Metrics

Various metrics are employed to assess the performance of ASR models, including Word Error Rate (WER), Cosine Similarity, and BLEU. These measurements play a crucial role in improving models and optimizing their accuracy.

WER is the most widely used metric for evaluating ASR systems. It quantifies transcription errors by computing the proportion of incorrect words relative to the total words in the reference text. Errors considered in WER include insertions, deletions, and substitutions within the generated transcription. The metric is formally expressed in Equation 1.

$$WER = \frac{I + R + D}{H + R + D} \tag{1}$$

where I is the number of inserted words, R denotes replaced words, D accounts for deleted words, and H corresponds to correctly transcribed words. Despite its extensive use, WER focuses strictly on word accuracy and does not account for contextual coherence.

Unlike WER, the BLEU metric [Papineni *et al.*, 2002] evaluates whether the transcription maintains the structure

and meaning of the original sentence. Originally developed for machine translation, BLEU has shown strong correlation with human evaluations. It relies on the precision of n-grams, comparing overlapping n-grams between the reference text T^* and the generated transcription T. The n-gram precision P_n is calculated by Equation 2.

$$P_n = \frac{|NG(n, T^*) \cap NG(n, T)|}{NG(n, T)} \tag{2}$$

BLEU is computed as the geometric mean of P_n for $n = 1, 2, 3, 4$, adjusted by a brevity penalty to discourage overly short transcriptions. The penalty factor, $bleu_{penalty}$, equals 1 when $|T| > |T^*|$ and is otherwise defined as $e^{(1-|T^*|/|T|)}$. The BLEU score is defined by Equation 3:

$$BLEU = \sqrt[4]{P_1 P_2 P_3 P_4} \times bleu_{penalty} \tag{3}$$

Cosine Similarity, another evaluation metric, measures the semantic similarity between two sentences by representing them in a vector space. This approach allows comparison beyond surface-level word accuracy. Word embeddings or sentence embeddings [Li *et al.*, 2020] (such as the Universal Sentence Encoder used in our experiments) can be applied to compute this metric. Cosine Similarity is defined as:

$$\begin{aligned} \cos(\mathbf{A}, \mathbf{B}) &= \frac{\mathbf{AB}}{\|\mathbf{A}\|\|\mathbf{B}\|} \\ &= \frac{\sum_{i=1}^{n} \mathbf{A}_i \mathbf{B}_i}{\sqrt{\sum_{i=1}^{n} (\mathbf{A}_i)^2} \sqrt{\sum_{i=1}^{n} (\mathbf{B}_i)^2}} \end{aligned} \tag{4}$$

where A and B are sentence vectors, with A_i and B_i representing their components, and $\|\mathbf{A}\|$ their Euclidean norms.

To assess the consistency of rankings given by different evaluators to LLM-generated transcriptions,we used Kendall's Tau Correlation Coefficient [Kendall, 1938], a non-parametric measure of ordinal association that quantifies the degree of agreement between two rankings. The coefficient τ is computed based on the number of concordant (c) and discordant (d) pairs, as shown in Equation 5. When ties are present in the data, a corrected version is applied (Equation 6), which accounts for the number of tied pairs in each ranking. The resulting τ value ranges from -1 (perfect disagreement) to $+1$ (perfect agreement), with 0 indicating no correlation.

$$\tau = \frac{c - d}{c + d} = \frac{S}{\binom{n}{2}} = \frac{2S}{n(n-1)} \tag{5}$$

$$\begin{aligned} \tau &= \frac{S}{\sqrt{n(n-1)/2 - T}\sqrt{n(n-1)/2 - U}} \\ T &= \sum_t t(t-1)/2 \\ U &= \sum_u u(u-1)/2 \end{aligned} \tag{6}$$

In addition to Kendall's Tau, Fleiss' Kappa [Fleiss, 1971] was employed to measure inter-annotator agreement when multiple evaluators assessed the quality of LLM-generated transcriptions. Unlike Cohen's Kappa [Cohen, 1960], which is limited to two evaluators, Fleiss' Kappa generalizes to scenarios involving three or more raters. It is computed from an $N \times k$ matrix, where each entry n_{ij} indicates how many raters assigned the i-th item to the j-th category, as shown in Equation 7. The final coefficient, given by Equation 8, compares the observed agreement (P_o) to the agreement expected by chance (P_e). The resulting κ ranges from -1 (complete disagreement) to $+1$ (perfect agreement), with values between 0.40–0.60 indicating moderate, 0.61–0.75 good, and above 0.75 excellent agreement.

$$\begin{aligned} p_j &= \frac{1}{Nn} \sum_{i=1}^{N} n_{ij}, \\ P_i &= \frac{1}{n(n-1)} \left(\sum_{j=1}^{k} n_{ij}^2 - n \right), \\ P_o &= \frac{1}{N} \sum_{i=1}^{N} P_i, \\ P_e &= \sum_{j=1}^{k} p_j^2, \end{aligned} \tag{7}$$

$$\kappa = \frac{P_o - P_e}{1 - P_e}. \tag{8}$$

4 Experimental Evaluation

4.1 Experimental setup

To guide this section, we have formulated the following key research questions. In what follows, we study each research question separately.

1. **RQ1.** What are some descriptive statistics from the testing data?
2. **RQ2.** Which are the pre-trained ASR models most effective for assisting with medical history?
3. **RQ3.** What decoding or post-processing strategies leveraging language models are most effective in improving the accuracy and reliability of ASR transcriptions?
4. **RQ4.** Which LLM is most effective in transferring the style of ASR transcriptions to align with structured patient record format?

4.2 Study on the results of RQ1

To explore the effects of automated transcription tools on cognitive load and efficiency in healthcare settings, this study analyzes recordings from specialists, comprising 382 audio clips of real patient interactions. Examination of these recordings provided key insights into their duration: the clips averaged 61.37 seconds, ranging from 3.24 to 104.30 seconds, with a standard deviation of 24.13 seconds.

According to Hapvida NotreDame Intermédica, medical consultations typically span around 15 minutes, with approximately half of this time — between 7 and 8 minutes — dedicated to manually inputting medical information into digital

systems. Given that the average recording duration in this dataset is significantly shorter than a full consultation, automated transcription tools such as ASR models could facilitate verification and re-recording when needed. This approach has the potential to reduce cognitive strain on healthcare professionals while improving workflow efficiency. Additionally, it may allow physicians to dedicate more time to direct patient interaction. These findings emphasize the importance of further research into the practical integration and effectiveness of automated transcription within medical environments.

4.3 Study on the results of RQ2

This study evaluates two pre-trained ASR models designed for Brazilian Portuguese: Wav2Vec2 PT[4] and Whisper[5]. The models were tested on a dataset of 606 transcribed audio recordings, which consists of recordings from both specialists and nonspecialists.

Using the WER metric, Whisper achieved an overall WER of 0.26, while Wav2Vec2 PT recorded 0.28. However, when analyzed separately, performance differences emerged between the two speaker groups. For nonspecialist recordings, Whisper obtained a WER of 0.37, a cosine similarity of 0.83, and a BLEU score of 0.43, whereas Wav2Vec2 PT achieved a WER of 0.24, a cosine similarity of 0.88, and a BLEU score of 0.55. In contrast, for specialist recordings, Whisper performed better with a WER of 0.19, a cosine similarity of 0.95, and a BLEU score of 0.67, compared to Wav2Vec2 PT, which obtained a WER of 0.30, a cosine similarity of 0.92, and a BLEU score of 0.51. These comparative results are summarized in Table 1.

Table 1. Performance Results for Whisper and Wav2Vec2 PT

Speaker Type	ASR Model	WER	BLEU (n-gram)	Cosine Similarity
Overall	Whisper	0.260	0.95	0.67
	Wav2Vec2 PT	0.280	0.88	0.55
Nonspecialist	Whisper	0.370	0.43	0.83
	Wav2Vec2 PT	0.240	0.55	0.88
Specialist	Whisper	0.190	0.67	0.95
	Wav2Vec2 PT	0.300	0.51	0.92

We conducted the Wilcoxon signed-rank test [Wilcoxon, 1992], a non-parametric statistical test that is preferable in this case because it does not assume normality in the data. The null hypothesis (H_0) posits that there is no difference in the WER performance between the Whisper and Wav2Vec2 PT models. The results of the test indicated significant differences between the models, with a p-value of 2.074e-05. These findings provide strong evidence to reject the null hypothesis, confirming that the models' performances are statistically distinct. Specifically, Whisper outperformed Wav2Vec2 PT, particularly with the specialist-recorded dataset.

Our evaluation also highlighted challenges faced by the ASR models, especially with medical terminology like "thyroid nodule (nódulo tireoidiano)," "sialosis (sialose)," and others listed in Table 2. We observed recurring errors in Brazilian Portuguese transcriptions due to phonetic similarities, such as between "SS" and "S" (ultrassonografia), "S" and "C" (sialose), and the silent "H" in words like "halitose." These observations emphasize the difficulties in accurately transcribing specialized medical vocabulary, revealing the limitations of widely used ASR models in handling complex healthcare terminology.

4.4 Study on the results of RQ3

We explored two strategies using language models to enhance ASR transcription corrections. The first involved training an n-gram model on nearly 5,000 medical reports from Hapvida NotreDame Intermédica, testing different values of $n \in \{3, 5, 7\}$. The approach integrated Wav2Vec2 PT with an n-gram model via KenLM. Despite variations in n, performance remained similar across models, as confirmed by Wilcoxon tests ($p > 0.05$, with p around 0.4), indicating no statistically significant differences. Given space limitations, we focus on presenting results for Wav2Vec2 PT+5-gram[6], which stood out in correcting transcription errors.

On the nonspecialist recorded dataset, the Wav2Vec2 PT+5-gram model achieved a WER of 0.18, a cosine similarity of 0.91, and a BLEU score of 0.65. However, performance declined on the specialist-recorded data, with WER rising to 0.37, while the cosine similarity remained at 0.91, and the BLEU score dropped to 0.42. Overall, the Wav2Vec2 PT+5-gram model obtained a WER of 0.30, a cosine similarity of 0.91, and a BLEU score of 0.50, slightly worse than the original model. These findings suggest that while the model adapts well to nonspecialist speech, it struggles more with specialist recordings, possibly due to variations in pronunciation, terminology, or speaking patterns.

To answer whether the post-processing step to correct spelling errors improves the accuracy and reliability of the ASR transcription, we evaluate the proposed model in multiple datasets. First, we evaluate the capacity of the model to identify spelling errors in general text, using a test dataset extracted from Wikipedia. Considering the micro-average of all classes, the model obtained an overall precision of 0.95, an overall recall of 0.94, and an overall F_1-score of 0.95. As such, we conclude that the model correctly detects the spelling errors we aimed to correct.

After that, we also evaluated the model in our medical domain datasets to assess whether its usage improves the reliability of the transcriptions. Table 3 shows the results for the nonspecialist dataset, with the best value of WER highlighted in bold. Overall, using the correction model improved the results, especially regarding scenario B, where correction is performed after normalization. We also compared the results without using correction to both scenarios where correction is applied. The Wilcoxon signed-rank test yielded a p-value of 0.11 for Whisper in scenario A, which was the highest among all comparisons. While this p-value does not indicate a statistically significant difference, it suggests weaker evidence of improvement for Whisper compared to the other ASR models, which all had p-values below 0.05.

Table 4 shows the results for the specialist dataset. The

[4] https://huggingface.co/jonatasgrosman/wav2vec2-xls-r-1b-portuguese

[5] https://huggingface.co/openai/whisper-large-v3

[6] https://huggingface.co/medtalkai/wav2vec_kenlm5

Table 2. Examples of medical histories generated by the ASR pre-trained models.

Model	Original	Transcription	WER	BLEU Score (n-gram)	Cosine Similarity
Wav2Vec2 PT	PT: paciente fez ultrassonografia que acusou nódulo tireoidiano. fez nova ultrassonografia com surgimento de um novo nódulo. EN: patient underwent ultrasound which accused of thyroid nodule. Performed another ultrasound with emergence of a new nodule.	paciente fez ultrasonografia que acusou o nódulo tireoidiano. fez nova ultrassonografia com o surgimento de um novo nódulo.	0.31	0.32	0.92
Whisper		paciente fez ultra sonografia que acusou o nóluo tiréoediano. fez nova ultra sonografia com o surgimento de um novo nóluo.	0.50	3.88e-78	0.91
Wav2Vec2 PT	PT: nega dor sialose halitose, pigarro ou outras queixas. EN: denies pain, sialosis, halitosis, throat clearing or other complaints.	nega dor cialose, alitose, pigarro ou outras queixas.	0.25	0.41	0.93
Whisper		mega doce se arose, alitosse, pigarro, ou outras queixas.	0.62	0.29	0.60

model improved the results of every ASR model, except for Whisper. As expected, the simulated ASR model had the smallest WER, with scenario B being the best. Leveraging the fact that the simulated ASR model only fails in the cases we aim to correct with the model, with the tokens perfectly aligned with ground truth, we obtained the token classification metrics. We found that the model also obtained good results in the medical domain, with an overall precision of 0.81, recall of 0.77, and F_1-score of 0.79.

Comparing actual ASR models, Whisper, with no correction, had the best result. This suggests Whisper already deals well with the spelling errors considered by the model and has other orthography challenges. Accordingly, analyzing the results, we found other spelling errors, like transcribing "hipotireoidismo" as "hipotireuidismo" and "submetido" as "submetida". Also, albeit rare, there are some wrong spelling error predictions, like "halitose" being incorrectly changed to "halitosse". In the nonspecialist dataset, Whisper, in scenario A, was also the only case where we found a p-value greater than 0.05 after conducting the Wilcoxon signed-rank test.

Table 3. Results on the nonspecialist dataset

ASR Model	Correction	WER	BLEU (n-gram)	Cosine Similarity
Whisper	No	0.3705	0.4265	0.8367
	Scenario A	0.3695	0.4283	0.8370
	Scenario B	**0.3690**	0.4293	0.8375
Wav2Vec2 PT	No	0.2413	0.5520	0.8848
	Scenario A	**0.2365**	0.5644	0.8859
	Scenario B	**0.2365**	0.5644	0.8860
Wav2Vec2 PT + 5-gram	No	0.1794	0.6544	0.9123
	Scenario A	**0.1770**	0.6585	0.9132
	Scenario B	**0.1770**	0.6585	0.9132

We conclude that using the correction model improves the accuracy and reliability of ASR transcriptions. However, other spelling challenges still need to be addressed and require a more comprehensive strategy for detection and correction.

Table 4. Results on the specialist dataset

ASR Model	Correction	WER	BLEU (n-gram)	Cosine Similarity
Simulated	No	0.0454	0.8904	0.9890
	Scenario A	0.0101	0.9756	0.9972
	Scenario B	**0.0064**	0.9845	0.9976
Whisper	No	**0.1981**	0.6769	0.9541
	Scenario A	0.1983	0.6767	0.9540
	Scenario B	0.1987	0.6756	0.9538
Wav2Vec2 PT	No	0.3095	0.5197	0.9281
	Scenario A	**0.3087**	0.5212	0.9283
	Scenario B	**0.3087**	0.5212	0.9284
Wav2Vec2 PT + 5-gram	No	0.3725	0.4238	0.9175
	Scenario A	**0.3711**	0.4261	0.9174
	Scenario B	**0.3711**	0.4261	0.9174

4.5 Study on the results of RQ4

The following pre-trained LLMs were used for the TST task of formatting medical anamneses in Portuguese: Llama-3.2-1B-Instruct[7], Mistral-Nemo-Instruct-2407[8], and Phi-3-mini-4k-instruct[9]. The evaluation process consisted of 70 transcriptions formatted by each model, which were then assessed by three healthcare students. The assessment focused on clarity, structural consistency, and adherence to medical formatting guidelines.

The Phi3 model demonstrated moderate performance, with an average score of 3.22, a median of 3, a mode of 3, and a standard deviation of 0.96, indicating low variability among evaluations. Scores ranged from 2.94 to 3.39, with standard deviations between 0.87 and 0.99. The Kappa coefficient was 0.179, reflecting low agreement among evaluators, while Kendall's correlation coefficients ranged from 0.29 to 0.52,

[7] https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct

[8] https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407

[9] https://huggingface.co/microsoft/Phi-3-mini-4k-instruct

suggesting moderate correlation. Main criticisms included a lack of precision, with partially correct or incomplete responses, and excessive use of generic language, limiting practical applicability.

The Llama model performed worse than Phi3, with an average score of 2.01, a median of 2, a mode of 1, and a standard deviation of 1.03, indicating greater variability and inconsistency in evaluations. The Kappa coefficient was 0.093, showing very low agreement, while Kendall's correlations ranged from 0.20 to 0.55. The model was criticized for frequently generating off-topic or irrelevant responses and for inconsistent formatting, making results harder to interpret.

The Mistral model achieved the best overall performance, with an average score of 4.00, a median of 4, a mode of 5, and a standard deviation of 0.93, reflecting consistently high evaluations. It had the highest Kappa coefficient (0.266), indicating moderate agreement, while Kendall's correlations ranged from 0.27 to 0.59. Despite its strong results, some evaluators noted excessive detail in responses, which, while thorough, sometimes reduced clarity and readability.

Example Input and Model Outputs

> **Input (Translated):** Comes for a neurological evaluation; his mother is concerned because he is very restless, does not learn, has poor interaction with other children, and is set to start school next month.

Model Outputs (Translated):

- **Phi-3:**

  Main complaint:

  - Reason for consultation: Excessive restlessness and difficulty learning interaction with other children.

- **Llama-3.2:**

  Main Complaint:

  - [Patient] is very restless and has difficulty learning poor interaction with other children.

- **Mistral-Nemo:**

  Main complaint:

  - Excessive restlessness
  - Difficulty interacting with other children
  - Concern about the start of school classes

5 Conclusion and Future Works

This paper addressed the critical need for accurate and efficient transcription of medical histories through the development and evaluation of an ASR tool. Given the absence of a dedicated Portuguese medical history database, we constructed a comprehensive benchmark consisting of 224 pairs of audio and text sourced from authentic medical history documents recorded by nonspecialists, as well as another one consisting of 382 longer pairs recorded by specialists. Our evaluation of established ASR models, including Wav2Vec2 and Whisper, revealed notable challenges, particularly in accurately transcribing specific medical terms. The intricate nature of healthcare-related vocabulary and the nuances of the Brazilian Portuguese language highlight the limitations inherent in widely used ASR models. Furthermore, we show that incorporating decoding and post-processing enhancement can improve transcription accuracy, especially for Wav2Vec2 PT, offering promising avenues for further research in medical transcription technology. Other improvements in decoding may be studied to find more consistent strategies, and the same goes for more comprehensive methods to detect and correct spelling errors.

In addition to ASR evaluations, we conducted a study on the results of the TST task for formatting medical anamneses in Portuguese using pre-trained LLMs. The models evaluated—Llama-3.2-1B-Instruct, Mistral-Nemo-Instruct-2407, and Phi-3-mini-4k-instruct—demonstrated varying performance levels. The Mistral model achieved the highest overall scores in clarity and adherence to medical formatting guidelines, whereas the Llama model performed the worst, often producing off-topic or inconsistently formatted responses. Phi-3 exhibited moderate performance but was limited by its tendency to use generic language. These results underscore the potential of LLMs in assisting medical documentation while highlighting the need for domain-specific optimization to improve accuracy and consistency.

A key limitation of this study was the inability to perform fine-tuning of ASR models due to technical challenges. The dataset contained long audio recordings, which conflicted with the constraints of models like Whisper, limited to 30-second inputs for fine-tuning. Additionally, our computational infrastructure was insufficient for such intensive tasks. Attempts to segment audios into smaller chunks were hindered by the low precision of forced alignment tools like Aeneas, making the process unfeasible. These issues prevented fine-tuning and highlight the need for better solutions to handle long audios and improve alignment accuracy.

In future research, we aim to fine-tune Mistral-7B-Instruct for the medical domain and explore the possibilities of multimodal large language models, such as Phi-4, to both transcribe audio and format medical case histories directly. We will also investigate advanced language models like GPT to enhance transcription accuracy. We also plan to investigate knowledge graphs and controlled vocabularies, comparing their performance to n-gram models or combining them. Additionally, we intend to address the technical limitations encountered by developing a robust audio segmentation process for the specialist dataset, enabling fine-tuning with long audio files. This will involve improving alignment precision and leveraging more efficient computational resources. These steps will be critical to advancing medical transcription technology and achieving more accurate ASR systems.

Acknowledgements

This research was conducted as part of the CEREIA project and was supported by multiple institutions and partners. We gratefully acknowledge the contributions from Hapvida NotreDame Intermédica, the Federal University of Ceará (UFC), and the São Paulo Research Foundation (FAPESP). We extend our gratitude to all the institutions,

partners, and funders involved in this project.

Funding

This research was supported by grant 2020/09706-7, São Paulo Research Foundation (FAPESP).

References

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In *NeurIPS*, pages 12449–12460.

Chiu, C.-C., Tripathi, A., Chou, K., Co, C., Jaitly, N., Jaunzeikare, D., Kannan, A., Nguyen, P., Sak, H., Sankar, A., *et al*. (2017). Speech recognition for medical conversations. *arXiv preprint arXiv:1711.07274*.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37 – 46.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Gonçalves, Y. T., Alves, J. V. B., Sá, B. A. D., da Silva, L. N., de Macedo, J. A. F., and da Silva, T. L. C. (2024). Speech recognition models in assisting medical history. In *Proceedings of the 2024 SBBD*.

Kar, S., Mishra, P., Lin, J., Woo, M.-J., Deas, N., Linduff, C., Niu, S., Yang, Y., McClendon, J., Smith, D. H., *et al*. (2021). Systematic evaluation and enhancement of speech recognition in operational medical environments. In *IJCNN*, pages 1–8.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, 30(1-2):81–93.

Lai, W., Hangya, V., and Fraser, A. (2024). Style-specific neurons for steering llms in text style transfer. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13427–13443.

Lee, T.-Y., Li, C.-C., Chou, K.-R., Chung, M.-H., Hsiao, S.-T., Guo, S.-L., Hung, L.-Y., and Wu, H.-T. (2023). Machine learning-based speech recognition system for nursing documentation–a pilot study. *IJMI*, 178:105213.

Li, B., Zhou, H., He, J., Wang, M., Yang, Y., and Li, L. (2020). On the sentence embeddings from pre-trained language models. In Webber, B., Cohn, T., He, Y., and Liu, Y., editors, *Proceedings of the EMNLP*, pages 9119–9130.

Li, J., Lavrukhin, V., Ginsburg, B., Leary, R., Kuchaiev, O., Cohen, J. M., Nguyen, H., and Gadde, R. T. (2019). Jasper: An End-to-End Convolutional Neural Acoustic Model. In *Proc. Interspeech 2019*, pages 71–75. ISCA. DOI: 10.21437/Interspeech.2019-1819.

Liu, Q., Qin, J., Ye, W., Mou, H., He, Y., and Wang, K. (2024). Adaptive prompt routing for arbitrary text style transfer with pre-trained language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18689–18697.

Mukherjee, S., Ojha, A. K., and Dušek, O. (2024). Are large language models actually good at text style transfer? *arXiv preprint arXiv:2406.05885*.

Paats, A., Alumäe, T., Meister, E., and Fridolin, I. (2018). Retrospective analysis of clinical performance of an estonian speech recognition system for radiology: effects of different acoustic and language models. *JDI*, 31(5):615–621.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th ACL*, page 311–318, USA. Association for Computational Linguistics. DOI: 10.3115/1073083.1073135.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *ICML*, pages 28492–28518.

Reddy, D. R. (1976). Speech recognition by machine: A review. *Proceedings of the IEEE*, 64(4):501–531.

Rubel Schneider, E. T., Andrioli de Souza, J. V., Knafou, J., Oliveira, L. E., Gumiel, Y. B., de Oliveira, L. F., Teodoro, D., Paraiso, E. C., Moro, C., *et al*. (2020). Biobertpt: a portuguese neural language model for clinical named entity recognition. In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*. 19 November 2020.

Schneider, S., Baevski, A., Collobert, R., and Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. In *Interspeech 2019*, pages 3465–3469.

Sullivan, P., Shibano, T., and Abdul-Mageed, M. (2022). Improving automatic speech recognition for non-native english with transfer learning and language model decoding. In *AANLSP*, pages 21–44.

Sunkara, M., Ronanki, S., Dixit, K., Bodapati, S., and Kirchhoff, K. (2020). Robust prediction of punctuation and truecasing for medical ASR. In Bhatia, P., Lin, S., Gangadharaiah, R., Wallace, B., Shafran, I., Shivade, C., Du, N., and Diab, M., editors, *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations*, pages 53–62, Online. Association for Computational Linguistics. DOI: 10.18653/v1/2020.nlpmc-1.8.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *NIPS*, pages 6000–6010.

Wilcoxon, F. (1992). Individual comparisons by ranking methods. In Kotz, S. and Johnson, N. L., editors, *Breakthroughs in Statistics: Methodology and Distribution*, pages 196–202. Springer New York, New York, NY.