Computers, Materials & Continua

Tech Science Press

Doi:10.32604/cmc.2025.061527

ARTICLE

Multilingual Text Summarization in Healthcare Using Pre-Trained Transformer-Based Language Models

Josua Käser[1], **Thomas Nagy**[1], **Patrick Stirnemann**[1] **and Thomas Hanne**[2,*]

[1]School of Business, University of Applied Sciences and Arts Northwestern Switzerland, Olten, 4600, Switzerland
[2]Institute for Information Systems, University of Applied Sciences and Arts Northwestern Switzerland, Olten, 4600, Switzerland
*Corresponding Author: Thomas Hanne. Email: thomas.hanne@fhnw.ch

Received: 26 November 2024; Accepted: 04 March 2025; Published: 26 March 2025

ABSTRACT: We analyze the suitability of existing pre-trained transformer-based language models (PLMs) for abstractive text summarization on German technical healthcare texts. The study focuses on the multilingual capabilities of these models and their ability to perform the task of abstractive text summarization in the healthcare field. The research hypothesis was that large language models could perform high-quality abstractive text summarization on German technical healthcare texts, even if the model is not specifically trained in that language. Through experiments, the research questions explore the performance of transformer language models in dealing with complex syntax constructs, the difference in performance between models trained in English and German, and the impact of translating the source text to English before conducting the summarization. We conducted an evaluation of four PLMs (GPT-3, a translation-based approach also utilizing GPT-3, a German language Model, and a domain-specific bio-medical model approach). The evaluation considered the informativeness using 3 types of metrics based on Recall-Oriented Understudy for Gisting Evaluation (ROUGE) and the quality of results which is manually evaluated considering 5 aspects. The results show that text summarization models could be used in the German healthcare domain and that domain-independent language models achieved the best results. The study proves that text summarization models can simplify the search for pre-existing German knowledge in various domains.

KEYWORDS: Text summarization; pre-trained transformer-based language models; large language models; technical healthcare texts; natural language processing

1 Introduction

Large, pre-trained transformer-based language models (PLMs) rely on deep learning and have become cutting-edge tools in natural language processing (NLP). These models aim at the language modeling objective and have shown outstanding capabilities in many tasks, particularly for texts written in English [1]. They thus enlarge the capabilities of previous NLP methods which can already support numerous use cases from individual document analysis up to social network analysis [2].

In this research project, we aim to analyze whether existing models are suitable means to perform the task of abstractive text summarization on German texts in the field of technical healthcare. Furthermore, we show how these models can be used to provide a suitable solution for the automatic summarization of texts and evaluate the quality of the results by conducting experiments.

The mentioned PLMs are mainly for English language applications, and multilingual adaptions may not support specific NLP functionalities for certain languages. Currently, little is known regarding the

multilingual capabilities since the pre-training of most models is performed on English texts [1]. Although it is known that PLMs may attain multilingual capabilities in this training setting, usually denoted as cross-language generalization, the quality and reliability of those capabilities remain to be tested [3].

Our research investigates the possibility of applying GPTs on non-English texts to perform abstract summarization, focusing on texts describing technical solutions in the healthcare field written in German. We focus on such technical texts in healthcare due to a general demand of providing summaries of technical texts. In addition, research on text summarization mostly focuses on general documents such as news articles, whereas particular domain-specific texts such as in the healthcare field are less investigated. Abstracts and summarizations of these texts would help researchers and domain experts review the literature and find relevant research more efficiently.

We consider the following hypothesis for guiding our research: Large language models can be used to perform a high-quality abstractive text summarization on texts describing technical solutions in the healthcare field written in German, although a used model is not specifically trained in the used language. Independently of the results regarding this hypothesis, the following research questions (RQ) will be answered to support a potential falsification or acceptance of the hypothesis:

- RQ1: Can transformer language models deal with complex syntax constructs of texts describing technical solutions in the healthcare field written in German?
- RQ2: Do transformer language models trained in English in the healthcare field perform better than transformer language models trained in German in the field of news articles when summarizing technical German healthcare texts?
- RQ3: How does the quality of abstractive summarization using transformer language models change when translating a source text to English before conducting the summarization?

The subsequent paper is organized as follows: In Section 2, we discuss related work. Our research methodology is explained in Section 3. Section 4 presents details about the developed artifact while its evaluation results including a discussion are shown in Section 5. Section 6 presents our conclusions.

2 Related Work

We conducted a literature review using a keyword-based search and forward/backward citation analysis. Searches were performed on Google Scholar and SpringerLink, employing the following keywords and their linguistic variants: "summarization NLP," "summarization healthcare," "summarization NLP German," "NLP abstract generation," "multilingual text summarization," "generative pre-trained transformer," "transformer-based language models," "evaluation of text summarization," and "NLP quality evaluation." This review yielded key findings, which are discussed below. While recognizing the importance of issues such as large language model security [4] and ethical considerations [5], these topics fall outside the scope of this study.

In the article of Min et al. [6], the authors surveyed recent work that uses large language models to solve NLP tasks via pre-training. The article references models such as Bidirectional Encoder Representations from Transformers (BERT) and GPT, which achieve state-of-the-art performances on many NLP tasks by leveraging the possibilities of deep neural networks. One of these tasks is text summarization, further described in an article by Zhang et al. [7]. The authors introduce two types of text summarization: extractive and abstractive. Extractive summarization selects salient sentences or phrases from a source text to generate the summary, while abstractive methods compose the summary by paraphrasing and restructuring sentences. Further development of deep neural networks improves extractive and abstractive summarization opportunities. Many works have a standard procedure for neural network-based methods containing three steps, as the authors mention in [8]. First, tokens are extracted from the source text into a continuous vector

representing semantic and syntactic information about the token. An encoder then processes this vector to output a fixed length vector called representations. Finally, the representations are provided to a generator to form an abstractive summarization. Furthermore, the authors introduce transformer models proposed in 2017 in the article "Attention is all you need" [9]. In this approach, an attention mechanism consisting of only self-attention and feedforward neural networks replaces the traditional usage of convolutional neural networks (CNNs) and recurrent neural networks (RNNs). Min et al. [6] reference GPT and BERT as models which achieve state-of-the-art performances in the field of NLP. Zaczynska et al. [10] explore the capability of the transformer model for the German language and show that models generally perform well but have limits regarding complex syntactic structures of the language.

The approach of multilingual pre-trained models, for example, multilingual BERT (M-BERT) [11], an adaption from BERT, enables users to build models where the target language differs from the main training data. Pires et al. [12] showed surprisingly good results in numerous languages in the models tested. However, they report significant differences in test results depending on the chosen language pair. Meanwhile, there are quite a number of multilingual, bilingual, or monolingual language models including German available. For instance, on the Huggingface platform we could identify seven models in total which support German.

The availability of data sets for text summarization in multiple languages is usually a problem, and constructing such a database is difficult [13]. Scialom et al. [14] presented a large-scale database for multilingual summarization. The database includes summary pairs in five different languages, and one of them is German. The database consists of more than 1.5 million articles and summaries gained through news articles. It would need to be tested if such databases can also be utilized for technical texts in healthcare [15].

The results of text summarization can be evaluated considering different dimensions. Gambhir et al. [16] categorized the methods into extrinsic and intrinsic evaluation measures. In the intrinsic measures, the text summarized by a machine is compared to a text summarized by a human. The human-made and machine-made text are compared in terms of the text's quality and informativeness. Recent evaluation techniques focus on automatic and semi-automatic methods to assess the informativeness of a text summary. Widely used automatic methods include ROUGE [17], which is the abbreviation of Recall-Oriented Understudy for Gisting Evaluation. The method considers the intersecting units of a text summary created by a machine to an ideal text summary made by a human based on n-gram, word sequences, and word pairs. Most published papers on text summarization do not rely on a single evaluation method to judge the informativeness of a text summary but rather show the results of different methods in one table [14,18,19]. Other methods, such as the factoid score [20] or the pyramid measure [21], were also introduced to assess informativeness in a semi-automatic manner. To assess the other dimension of the intrinsic category, the quality of the summarization outcome needs to be tested by checking the grammar, structure, coherence, focus, and non-redundancy [18]. The extrinsic evaluation is executed by assessing the relevance and comprehensiveness of the machine-made summary with the involvement of humans [16].

Rohil et al. [22] argue that while authors are summarizing literature in other fields, this is usually not the case for texts in the medical or healthcare area. However, multiple studies have already been conducted in this field. Gigioli et al. [23] use deep reinforcement learning to produce summaries of one sentence length from texts in the medical domain. A model proposed by Gayathri et al. [24] uses a domain-specific vocabulary thesaurus (MeSH) to rank sentences by importance. Finally, Moradi et al. [25] use the before-mentioned BERT model and apply it to texts in the medical and healthcare domain. Zesch et al. [26] discuss the processing of German medical text with NLP techniques. However, their focus is mainly on texts such as patient documentation in form of clinical notes and specific tasks of text summarization are not discussed. Another study focusing on the summarization of electronic medical records in English has been provided by Bi et al. [27]. A recent review of the usage of pre-trained language models in medicine in general has been

provided by Luo et al. [28]. Recent results on German text summarization (not focusing on medical texts) are provided by Schubiger [29] and a general recent survey on text summarization is given by Zhang et al. [30].

Table 1 presents an overview of the discussed results regarding their contribution to areas such as text summarization, German or multilingual application, and the focus on the medical domain. While all previous studies have been successful in their ways, there is a clear research gap regarding the suitability of large language models applied to technical texts regarding healthcare written in German.

Table 1: Overview of existing research regarding text summarization, German or multilingual application, and the focus on the medical domain

	Text summarization	German	Multilingual	Medical texts
[27]	X			X
[11]			X	
[16]	X			
[24]	X			X
[23]	X			X
[17]	X			
[18]	X			
[28]				X
[19]	X			
[25]	X			X
[21]	X			
[22]	X			X
[12]			X	
[20]	X			
[29]	X	X		
[14]	X			
[10]		X		
[26]		X		X
[7]	X			
[30]	X		X	
Our study	X	X	X	X

3 Research Design

The first subsection describes the research methods considered as well as the research strategy. The subsequent sections cover the technological concepts considered for the development of the artifact, the data collection approach, and the analysis and evaluation methods.

3.1 Research Method and Strategy

We use an experimental research method for the study of multilingual text summarization. This research method addresses the study purpose because the hypothesis and the associated research questions can be answered by the evaluation and following comparison of different language models described in the research questions in Section 1.

Due to the nature of the hypothesis, an artifact body of code is needed. We will develop this artifact during the study and denote it as "prototype" in the following sections. This prototype will be used to conduct

experiments on technical healthcare texts in German containing no summarization or abstract. Fig. 1 shows the applications scenario using the two approaches for answering research question 3 (RQ3), the direct application of a respective model and the prior automatic translation to English, and a subsequent translation of the results (generated abstracts) back to the target language. The input for both approaches will be the same, i.e., German language texts, and the final results are summaries in German, so that both approaches are comparable regarding the quality of results produced.

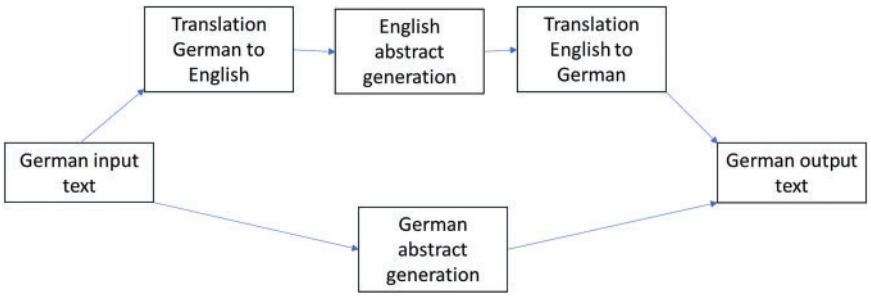

Figure 1: Application scenario of two approaches to be compared

3.2 Relevant Technological Concepts

The following subsections explore technological concepts that support the development of a prototype further used as a test environment to evaluate different language models.

3.2.1 Natural Language Processing and Text Summarization

According to Liddy [31], Natural Language Processing is based on a set of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis to achieve human-like language processing for a range of tasks or applications.

Munot et al. [32] describe various text summarization methods and the differences in the approaches. Generally, they define text summarization as reducing the original document's size while preserving its information content, and its summary is less than half of the main text. Furthermore, as mentioned in the Literature Review, there are generally two different methods in automated text summarization, extractive and abstractive.

- Extractive summarization is described by Liu [33] as follows. The method analyzes a document or several documents on a sentence and word-based approach, where the algorithm indicates whether a sentence should be included in the summary, thereby it is assumed that sentences that represent the essential content of a document should be incorporated in the summary.
- Abstractive summarization is described by Moratanch et al. [34] as follows: The method performs summarization by understanding the original text with the help of linguistic methods to understand and examine the text. The objective of the summarization is to create a generalized summary, which conveys information in a precise way that generally requires advanced language generation and compression techniques. This results in a summary created with entirely new sentences to deliver the most relevant information.

For our study, the method of abstractive summarization is used to analyze the capability of different language models.

3.2.2 Development Environment

The first choice of a programming language when it comes to text data manipulation is Python which offers a number of benefits. As a significant number of open-source NLP libraries are available in Python, and other machine learning libraries provide Python APIs, we decided to use Python as the primary programming language for this project. A particular set of software programs is required to work with human language data in Python. A well-known platform called the Natural Language Toolkit (NLTK) offers a variety of corpora, lexical resources, and libraries for text categorization, parsing, tokenization, and semantic reasoning. Furthermore, NLTK includes some basic metrics for the evaluation of NLP tasks [35] which are useful for our analysis of text summarization capabilities.

3.2.3 Language Models

In almost every NLP task, transformer-based pretrained language models have succeeded remarkably well during recent years. These models are based on a transformer neural network trained in a self-supervised scenario using large volumes of text data. Transformer-based PLMs achieve universal language representations and transfer learning capabilities enable them to succeed in NLP tasks which were not part of the training scenario. For example, NLP researchers created models like BERT and GPT-3 by pre-training them on massive amounts of unlabeled text using self-supervised learning, which was motivated by the success of pre-trained image models [36].

GPT-3 (Generative Pre-trained Transformer 3) is a state-of-the-art language model developed by OpenAI. It uses a deep neural network with 175 billion parameters, making it one of the largest models of its kind. In addition, the model is pre-trained on a massive dataset of diverse text, allowing it to generate human-like text when finetuned on specific tasks. The Text-Davinci-003 is a specific version of the GPT-3 model with the highest API access level.

GPT-3 uses a transformer architecture, which uses self-attention mechanisms to weigh the importance of different parts of the input, allowing the model to focus on the most relevant information [9]. GPT-3 is trained using a variant of the transformer architecture called the Transformer-XL, which was introduced by Dai et al. [37]. The Transformer-XL allows the model to learn patterns in text that span longer distances, making it more effective at understanding the meaning and context of the text. The GPT-3 model is finetuned on specific tasks using the transfer learning method, which allows the model to adapt to new tasks using the knowledge it has learned from pre-training. This approach is highly effective in various natural language processing tasks and has been widely used in recent years in scientific literature.

BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained transformer-based language model developed by Google that has achieved state-of-the-art performance on a wide range of natural language understanding tasks. The model is trained using masked language modeling, in which some of the words in the input are replaced with a special token (e.g., [MASK]), and the model is trained to predict the original token based on the context. The BERT model is a bidirectional model, which means it considers the context before and after a given token, as opposed to traditional models that only consider the context before the token. This bidirectional approach allows BERT to understand the text's meaning and context better. The model is pre-trained on a large corpus of text and finetuned on specific tasks such as sentiment analysis, question answering, and named entity recognition. Finetuning is done by adding a small task-specific layer on top of the pre-trained BERT model and training it on the task-specific dataset [11].

T5 (Text-to-Text Transfer Transformer) is a state-of-the-art natural language processing (NLP) model developed by Google Research. It is a variant of the Transformer architecture designed to generate text from text inputs. The model is trained on various tasks, such as text classification, summarization, translation,

and question answering. The model works by encoding an input sequence of text and generating an output sequence of text. First, the model is trained using a masked language modeling objective, where a portion of the input tokens are masked, and the model is trained to predict the masked tokens given the context. During inference, T5 takes an input text string and adds a special token to indicate the start and end of the input text. It then uses its pre-trained encoder to convert the input text into a sequence of continuous representations, known as embeddings. The decoder then uses these embeddings to generate the output text. One of the fundamental features of T5 is its ability to perform any NLP task by simply modifying the input and output text strings without any task-specific modification to the model architecture. This is achieved by providing the task information as a text string that is appended to the input text, which the model uses to determine the desired output. T5's large size and pre-training on a diverse range of tasks enables it to have a high level of generalization, allowing it to perform well on a wide range of NLP tasks without the need for task-specific finetuning [38].

3.3 Data Collection and Evaluation Approaches

Different approaches are possible for creating the dataset serving as the input of the system. Webscraping and webcrawling are techniques to automatically search for unstructured data on the web and store them in a structured format. They are unable to ascertain whether the collected data is useful as input to the proposed prototype solution. This is due to the specificity of the input needed. No single database or set of parameters exists for technical texts in the healthcare domain; thus, this approach is not viable for application in this case. Because the prototype should only serve as a proof of concept, a limited sample of 16 documents is collected manually from the healthcare magazine Aphasie Suisse [39], which publishes recent advances and studies in the field of aphasia. To aid the evaluation process described in the next paragraphs, these texts already contain abstracts that can be used to evaluate the quality of the results.

As described in the literature review, the taxonomy laid out by Gambhir et al. [16] suggests evaluating text summarization models based on intrinsic and extrinsic measurements. However, our research questions primarily focus on comparing different models, and we do not attempt to use the output to perform subsequential tasks such as question answering with the generated summary. Therefore, we focus this paper on an intrinsic evaluation and will assess the two categories of informativeness and quality.

3.3.1 Informativeness

To assess the informativeness, we will rely on the well-established ROUGE methods [17]. The ROUGE methods take an automatically summarized text and compare it to a reference summary. This comparison can be made by considering the recall or the precision measure. The recall measure compares the number of matching words in an automatically summarized text to the total number of words in the reference summary. The precision measure compares the number of overlapping words to the total number of words in the automated text summary to ensure that only relevant words are included in the summary. The F-Measure (or F1 score) combines recall and precision by multiplying these two values and dividing the result by their sum, and multiplying by two.

$$F1\,score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The ROUGE-N method assesses a certain number of grams. ROUGE-1 assesses unigrams (single words), and ROUGE-2 bigrams where two words are taken as a comparison. The ROUGE-L method measures the longest common sequence (LCS) of words. The advantage is that the n-gram length does not need to be defined, as it automatically includes the longest common sequence. There are many more ROUGE

variants available, but Lin et al. [40] show encouraging results for the ROUGE-L method to assess the quality of automatic machine translation. Therefore, we will use a combination of the F-Measure of ROUGE-1, ROUGE-2, and ROUGE-L to assess the informativeness aspect of our German text summarization model.

3.3.2 Quality

The conferences of DUC (Document Understanding Conference), which became later a part of TAC (Text Analysis Conference), suggest that linguistic quality is best assessed by a human reviewer rating the aspects of non-redundancy, focus, grammar, referential clarity, and structure and coherence on a scale from 1–5 [18].

- Grammaticality: The text summary should not consist of grammatical errors such as capitalization errors, missing words, or fragments that would hinder the reader's understanding of the sentence.
- Non-redundancy: The summary should be free of needless repetitions, such as whole sentences or facts that only need to be mentioned once.
- Referential clarity: It should be clear what pronouns and noun expressions refer to. If an object or person is mentioned, their significance to the story should be easily identified.
- Focus: The summary should only include focused information that is important in the whole context of the text.
- Structure and coherence: A straightforward design should exist, as well as a logical order of sentences in summary. Sentences should be logically connected.

There are also some automatic attempts made by Pitler et al. [41], which only show similar performance to a human reviewer when performing several statistical models in combination, which seems not to provide enough added value to be used in the framework of this paper, so we refer to the five linguistic questions of the DUC/TAC conference. The assessment is done by the authors using a 5-point Likert scale.

4 Development and Artefact

4.1 Implementation Process

Before implementing the prototype, we defined the necessary steps to conduct experiments to answer the research questions. The following list describes the implemented process steps to generate summaries:

1. Read input text.
2. Pre-process the input text based on the two substeps:
   2.1. Optional: Translate text to English.
   2.2. Split text into subtexts (due to input length limits of PLMs).
3. Summarize each subtext using a pre-trained language model.
4. Merge subtext into the final summary (concatenation of subtexts).
   4.1. Optional: Translate the summary back to German.
5. Calculate evaluation metrics.

We used Python 3.10.9 programming language to develop the prototype. For simplicity reasons, the resulting prototype artifact does not entail a graphic user interface. The technology used, as well as the mechanics, can be summarized as follows. The prototype is based on several libraries, packages, and bundles documented in a requirements file. It should be noted that some models of the prototype use the API provided by OpenAI, which could generate costs [42]. The API is used to implement translation and summarization with GPT-3. Furthermore, Hugging Face transformers library is used for other language

models such as BERT and a Bio-Medical Model [43]. We did not consider additional efforts to optimize the runtime as the prototype is used for research purposes only.

Generally, the provided code summarizes all txt-files included in the input folder and places them in the output folder. An independent script is used to create the evaluation files, which considers the generated and corresponding target summaries. The evaluation metrics are calculated by using the rouge-score Python library and functionalities of the NLTK Python library.

4.2 Implemented Approaches

The following list provides an overview of the four implemented approaches to perform abstractive text summarization:

1. GPT-3 approach (GPT-3): This approach uses the GPT-3 language model described in Section 3.2.3. to generate the summary. The implementation accesses the OpenAI API to summarize the German input text.
2. Translate approach (T): This approach utilizes the same API and language model as the GPT-3 approach. Before the German input text is sent to the API, the input text is translated into English. After performing the summarization task, the text is translated back into German.
3. German language Model approach (GM): This approach uses an implementation of the BERT language model described in Section 3.2.3. to generate the summary. This multilingual model was trained using data from five languages. It is accessed via Hugging Face transformers library, and the full description can be found at https://huggingface.co/mrm8488/bert2bert_shared-german-finetuned-summarization (accessed on 03 March 2025).
4. Bio-Medical language model approach (BM): This approach uses an implementation of the T5 language model described in Section 3.2.3. to generate the summary. The model is specifically trained to perform abstractive summarization on biomedical research papers. The finetuning of the model was performed on the Scitldr [44] dataset. The model is accessed via Hugging Face transformers library, a description of the model can be found at https://huggingface.co/spaces/Blaise-g/summarize-biomedical-papers-long-summary-or-tldr (accessed on 03 March 2025).

The considered models are used for summarization with default settings and most of them do not provide a further specification of the length of the summary. In our case (and probably for the evaluation of text summarization in general) it would be most useful to target a summary length similar to already available summaries. Only for the German BERT model, we could specify a minimum length of 150 and a maximum length of 200 tokens which appears similar to the target lengths.

5 Evaluation

5.1 Evaluation of Informativeness

Fig. 2 presents a comparison of the performance of four distinct models; Bio-Medical (BM), German/Multilingual based on BERT (GM), GPT-3, and a model that utilized a two-step process involving translation to English and subsequent GPT-3 summary generation followed by translation back to German (T). The comparison was conducted using three key metrics: ROUGE-1 (R1) F-Measure, ROUGE-2 (R2) F-Measure, and ROUGE L (RL) F-Measure. The results of the R1 F-Measure, which evaluates the number of single words overlapping between the generated summary and the ideal summary, indicate that the GPT-3 model achieved the highest average score of 0.322, followed by the BM model with an average score of 0.256, the GM model with 0.231, and the T model with 0.221. Additionally, the GPT-3 model showed the highest average score for the R2 F-Measure, evaluating two consecutive words, with a value of 0.092. The ROUGE

L F-Measure, which evaluates the longest common sequence, also demonstrated the GPT-3 model as the leader with a score of 0.152.

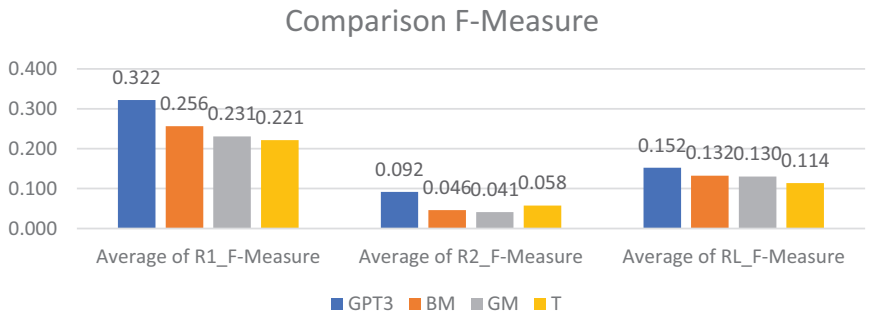

Figure 2: Comparison of the results using the F-Measure

The GPT-3 model required 320 words on average to create a summary, which is significantly lower than the T model, which utilized 612 words on average. However, it is also substantially higher than the GM model, which required 112 words, and the BM model, which utilized 85 words. The differences in lengths of the generated summaries significantly limit the possibilities to compare the results of the considered models.

As depicted in Figs. 3 and 4, the length of the summary has a substantial impact on the R1 precision and the R1 recall. Longer generated summaries are more likely to match the words of the ideal summary, thus increasing the recall value, while excessive words in the generated summary can decrease the precision value. Finally, it is worth mentioning that although the GPT-3 model uses fewer words compared to the T model (see Table 2), it still achieves similar recall scores. This is because the translation process used by the T model often adds additional, non-value-adding words to the summary.

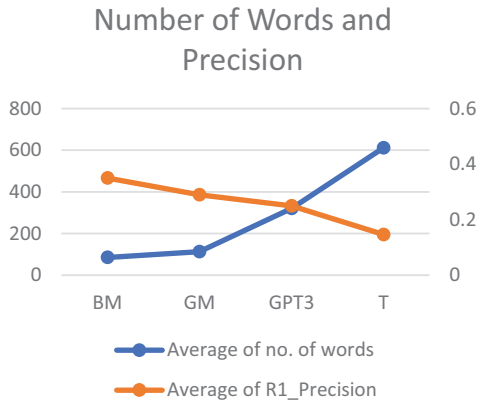

Figure 3: Number of words and precision

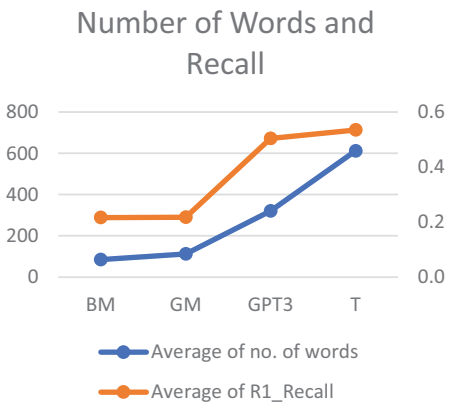

Figure 4: Number of words and recall

Table 2: Recall measure comparison for GPT-3 and T

Language model	Average of no. of words	Average of R1_Recall	Average of R2_Recall	Average of RL_Recall
GPT-3	320	0.504	0.140	0.240
T	612	0.535	0.137	0.281

5.2 Evaluation of Quality

In addition to the automated assessment based on ROUGE scores, we conducted a manual assessment of five quality aspects for the results of the model as described in Section 3.3.2. The results of the manual quality assessment align with those of the automated assessment. As shown in Table 3, the GPT-3 language model achieved the highest score, with an average of 3.90 points out of a maximum of 5. The T model and the BM model followed closely with scores of 3.73 and 3.68 points, respectively. The only model that performed significantly lower was the GM model, with a score of only 1.85 points.

Table 3: Average quality scores (see Fig. 5)

Model	Average of scores
GPT-3	3.90
T	3.73
BM	3.68
GM	1.85

As shown in Fig. 5, the GPT-3 model did not exhibit the highest performance in any of the five categories, but it did not present a specific weakness in any of the dimensions, unlike the other models.

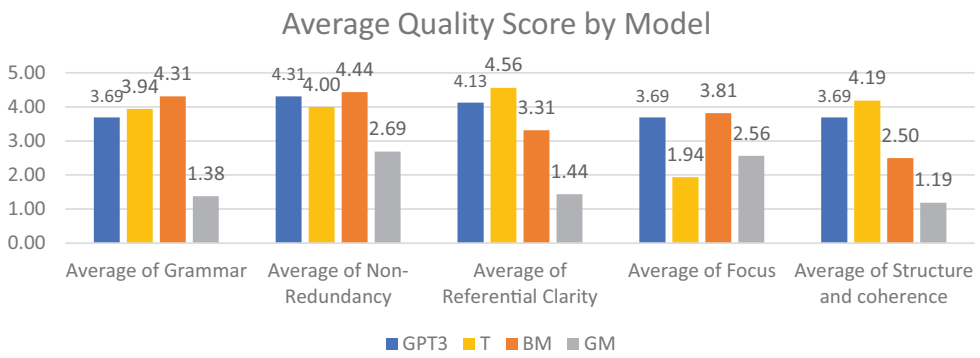

Figure 5: The five quality scores of the considered models

The T model performed poorly in the Focus category, while the BM model struggled with Structure and Coherence. On the other hand, the German Multilingual Model (GM) based on the MLSUM dataset underperformed across all categories. Despite selecting relevant words, resulting in an average score for informativeness based on the ROUGE scores, the sentences generated by the GM model often lacked coherence for a human reader. Fig. 6 provides an example of such a generated text. This highlights the importance of automated ROUGE scores and manual assessment by a human expert in determining the quality of a text summary, as manual assessment offers a more profound understanding beyond the ROUGE score.

```
In manchen Sprechangst und Aphasie ist es die Sprechstunde, die die Sprache in ihrer Lebensqualität stark verändert.
Eine Übersicht über die wichtigsten Diagnosen, wie sich Patienten schützen können - und was es sich lohnt,
die Diagnose zu beantworten und was sie in der Medizin und Psychotherapie zu lesen sind.
```

Figure 6: Example 1 of a GM summary

Other outputs of the GM model, for example, shown in Fig. 7, were distorted and would need further investigation if finetuning the model would lead to better results.

```
Bonus für Patienten, Diabetes und Diabetes, Diabetes, Infusus und Psycho -xie. Bonus. Bonus, Medikamente, Psychotherapien, Diabetes -
und psychischer Rückschlagtage zu testen, Psychotesten, Diabetes oder Kranken - und Patienten - und was Sie auf dem Weg zum Besseren?
```

Figure 7: Example 2 of a GM summary

The Biomedical Model with Translation demonstrated the ability to generate coherent sentences without grammatical errors. Despite producing concise summaries with an average of 14 words per sentence, the shortest among all the models, it struggled to connect the sentences to form a cohesive narrative. Many sentences started with similar phrases, as illustrated in Fig. 8, where two sentences start with “Wir lernen,” but lack a connection between them. As a result, the model received a low score in the Structure and Coherence category.

In most categories, the Translated GPT-3 Model performed similarly or even better than the original GPT-3 Model. However, it generated excessively lengthy summaries, impacting its focus. Additionally, as depicted in Fig. 9, some words were not correctly translated back to German and remained in English, leading to potential linguistic errors.

Wir lernen, wie man digitale Therapiematerialien für Kinder und Erwachsenentherapie entwirft.

Wir lernen, digitale Geräte und elektronische Therapiematerialien in der Sprachtherapie zu nutzen.

Figure 8: Example of BM model summary

Aphasiespezifische Faktoren wie die Art und Schwere der aphasischen Symptome und die Phase der Störung können auch die Redeangst der Betroffenen beeinflussen.
Die Symptome der Redeangst manifestieren sich auf drei Ebenen: kognitiv, Verhaltens- und körperlich.
Darüber hinaus können langfristige psychosoziale Folgen aus der Störung resultieren. available to cope with the situation.

Figure 9: Example of T model summary

The untranslated GPT-3 Model generated comprehensible summaries with a correct grammatical structure, and effectively handled specialized medical terms. However, its challenge was the limitation of tokens, which caused the summaries to end abruptly, as demonstrated in Fig. 10. During the quality evaluation, these incomplete sentences were considered. However, further investigation is necessary to address this token limitation issue, which could result in higher scores in the quality assessment.

Es gibt viele Regeln, die beachtet werden müssen, um eine gute Benutzererfahrung zu gewährleisten, wie z.B. eine klare und gut lesbare Schrift,
eine ansprechende Gestaltung, eine effektive Dokumentation und eine ansprechende Gamifikation. Es ist auch :
Für eine effektive Sprachtherapie ist es wichtig, dass Therapeuten mit neuen Technologien vertraut sind und grundlegende Kenntnisse über das En

Figure 10: Example of GPT-3 model summary

5.3 Discussion

In this subsection, we discuss how this study has led to answering the research questions and proving the hypothesis formulated in Section 1 to be true, thus closing the research gap.

RQ1: Can transformer language models deal with complex syntax constructs of texts describing technical solutions in the healthcare field written in German? The results evaluating the summaries made by the different models have shown (both for the ROUGE based automatic evaluations and for manual assessment) that it is entirely possible for these models to deal with the complex technical input texts provided despite clear potential for further improvement.

RQ2: Do transformer language models trained in English in the healthcare field perform better than transformer language models trained in German in the field of news articles when summarizing technical German healthcare texts? This research question can be answered by comparing the Bio-Medical language model with the German language model. The German language model underperformed in terms of manual quality assessment compared to all other models. However, since only one German-trained model has been tested, it is not possible to prove conclusively that German models are worse. On the other hand, the Bio-Medical language model used had trouble to form a cohesive narrative. With these caveats in mind, this research question can be answered with “Yes” even though both models are performing worse than the other tested models. As multilingual or German language models lag significantly behind state-of-the-art English language models, future progress may lead to another answer to this question.

RQ3: How does the quality of abstractive summarization using transformer language models change when translating a source text to English before conducting the summarization? Evaluating the proposed translated model has shown that using a translator-based approach is very promising. Especially, for two categories of the manual assessment, the T model with translation achieved the best results. However, many nuances and details get lost by translating such technical texts as used in this work.

6 Conclusions

With the answering of the research questions based on the experiments conducted, we believe the hypothesis to be proven true despite limitations of our study such as regarding the limited range of considered models, limitations of finetuning them for the specific application (e.g., regarding the length of summaries), or concerning the limited amount of sample data. Even with these limitations, it can be clearly stated that text summarization models can be used in the healthcare domain even if the text produced is German. The performance of the models not specifically trained in one domain shows great promise of use in other knowledge domains, as the results proved that domain-specific knowledge is of limited use when summarizing a text. However, there is still quite a significant potential for further improvement. This technology may be used by domain specialists searching for knowledge databases without author-provided abstracts. Due to the limited number of models tested and the single domain analyzed, it is impossible to conclusively prove that this is possible in every domain or that the models used to evaluate this domain would perform equally in other areas. As shown in Section 3.3, certain models produced strange results, likely due to coding errors. Whether implementing these models in lesser-used languages than German produces the same quality of results would have to be studied in further detail. In addition, we suggest using a broader range of large language models in future research such as the most recent versions of GPT, Gemini, or Llama. In addition, in subsequent studies, we will explore further techniques such as retrieval augmented generation in medical use cases employing large language models.

In general, we assume that further significant progress will be made in large language models, not only in general but also regarding multilingual or specific non-English models, in the coming years. Therefore, further studies are required. Moreover, such studies should also cover a wider range of suitable test data (such as technical or academic documents related to healthcare and medical topics). Specific focus should be given to the dependence of summarization quality on the level of technical language and the complexity of grammatical structure used in the sample texts.

In conclusion, this work delivered proof that text summarization models can be used in German healthcare texts' very specific and technical domain. As domain-independent language models like GPT-3 achieved the best results, this technology has the potential to be helpful to many experts in practice and academia by simplifying the search for pre-existing German knowledge in their domains. Indeed, we further experimented with this approach and provided a respective solution for a platform supporting digital maturity assessment and further digitalization in the healthcare field [45]. Experiments with this platform indicated that automatically generated summaries are understandable by healthcare professionals who considered them very useful for providing fast insights into the contents of respective documents. During this project, we also considered using alternatives to well-known models such as those by OpenAI for cost reasons or for confidentiality considerations which appeared possible with acceptable computational effort due to reduced model complexities but still with acceptable quality of the results.

Acknowledgement: None.

Funding Statement: No funding was received for conducting this study.

Author Contributions: The authors confirm contribution to the paper as follows: methodology, investigation, writing—original draft preparation: Josua Käser, Thomas Nagy, Patrick Stirnemann; validation: all authors; writing—review and editing, supervision: Thomas Hanne. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: Detailed data from the evaluation (ROUGE scores and manual assessment) is available from the authors on request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

1. Armengol-Estapé J, de Gibert Bonet O, Melero M. On the multilingual capabilities of very large-scale English language models. arXiv:2108.13349. 2021.
2. Cauteruccio F, Corradini E, Terracina G, Ursino D, Virgili L. Extraction and analysis of text patterns from NSFW adult content in Reddit. Data Knowl Eng. 2022;138(8):101979. doi:10.1016/j.datak.2022.101979.
3. Zhang X, Li S, Hauer B, Shi N, Kondrak G. Don't trust ChatGPT when your question is not in English: a study of multilingual abilities and types of LLMs. arXiv:2305.16339. 2023.
4. Luo H, Luo J, Vasilakos AV. BC4LLM: a perspective of trusted artificial intelligence when blockchain meets large language models. Neurocomputing. 2024;599(6):128089. doi:10.1016/j.neucom.2024.128089.
5. Jiao J, Afroogh S, Xu Y, Phillips C. Navigating LLM ethics: advancements, challenges, and future directions. arXiv:2406.18841. 2024.
6. Min B, Ross H, Sulem E, Ben Veyseh AP, Nguyen TH, Sainz O, et al. Recent advances in natural language processing via large pre-trained language models: a survey. arXiv:2111.01243. 2021.
7. Zhang H, Cai J, Xu J, Wang J. Pretraining-based natural language generation for text summarization. In: Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL); 2019; Hong Kong, China. p. 789–97. doi:10.18653/v1/k19-1074.
8. Wang G, Smetannikov I, Man T. Survey on automatic text summarization and transformer models applicability. In: 2020 International Conference on Control, Robotics and Intelligent System; 2020; Xiamen, China. p. 176–84. doi:10.1145/3437802.3437832.
9. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. arXiv:1706.03762. 2017.
10. Zaczynska K, Feldhus N, Schwarzenberg R, Gabryszak A, Möller S. Evaluating German transformer language models with syntactic agreement tests. arXiv:2007.03765. 2020.
11. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2019; Kerrville, TX, USA: The Association for Computational Linguistics. p. 4171–86. doi:10.18653/v1/N19-1423.
12. Pires T, Schlinger E, Garrette D. How multilingual is multilingual BERT? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019; Florence, Italy. p. 4996–5001. doi:10.18653/v1/p19-1493.
13. Parida S, Motlicek P. Abstract text summarization: a low resource challenge. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); 2019; Hong Kong, China. p. 5993–97. doi:10.18653/v1/d19-1616.
14. Scialom T, Dray P-A, Lamprier S, Piwowarski B, Staiano J. MLSUM: the multilingual summarization corpus. arXiv:2004.14900. 2020.
15. Frefel D. Summarization corpora of wikipedia articles. In: Proceedings of the Twelfth Language Resources and Evaluation Conference; 2020; Marseille, France. p. 6651–55.
16. Gambhir M, Gupta V. Recent automatic text summarization techniques: a survey. Artif Intell Rev. 2017;47(1):1–66. doi:10.1007/s10462-016-9475-9.
17. Lin C-Y. ROUGE: a package for automatic evaluation of summaries. In: Text summarization branches out. Kerrville, TX, USA: The Association for Computational Linguistics; 2004. p. 74–81.
18. Lloret E, Plaza L, Aker A. The challenging task of summary evaluation: an overview. Lang Resour Eval. 2018;52(1):101–48. doi:10.1007/s10579-017-9399-2.
19. Mani I, Klein G, House D, Hirschman L, Firmin T, Sundheim B. SUMMAC: a text summarization evaluation. Nat Lang Eng. 2002;8(1):43–68. doi:10.1017/S1351324901002741.

20. Teufel S, van Halteren H. Evaluating information content by factoid analysis: human annotation and stability. In: Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing; 2004; Kerrville, TX, USA: The Association for Computational Linguistics. p. 419–26.
21. Nenkova A, Passonneau R. Evaluating content selection in summarization: the pyramid method. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL; 2004; Kerrville, TX, USA: The Association for Computational Linguistics. p. 145–52.
22. Rohil MK, Magotra V. An exploratory study of automatic text summarization in biomedical and healthcare domain. Healthc Anal. 2022;2(2):100058. doi:10.1016/j.health.2022.100058.
23. Gigioli P, Sagar N, Voyles J, Rao A. Domain-aware abstractive text summarization for medical documents. In: 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM); 2018 Dec 3–6; Madrid, Spain: IEEE; 2018. p. 1155–62. doi:10.1109/BIBM.2018.8621457.
24. Gayathri P, Jaisankar N. Towards an efficient approach for automatic medical document summarization. Cybern Inf Technol. 2015;15(4):78–91. doi:10.1515/cait-2015-0056.
25. Moradi M, Samwald M. Clustering of deep contextualized representations for summarization of biomedical texts. arXiv:1908.02286. 2019.
26. Zesch T, Bewersdorff J. German medical natural language processing—a data-centric survey. Applicat Med Manufact. 2022; 137–45.
27. Bi B, Liu L, Perez-Concha O. Adapting large language models for automated summarisation of electronic medical records in clinical coding. Stud Health Technol Inform. 2024;318:24–9. doi:10.3233/SHTI240886.
28. Luo X, Deng Z, Yang B, Luo MY. Pre-trained language models in medicine: a survey. Artif Intell Med. 2024;154(10):102904. doi:10.1016/j.artmed.2024.102904.
29. Schubiger R. German summarization with large language models [master's thesis]. Zurich, Switzerland: ETH Zurich; 2024.
30. Zhang H, Yu PS, Zhang J. A systematic survey of text summarization: from statistical methods to large language models. arXiv:2406.11289. 2024.
31. Liddy ED. Natural language processing [Internet]. 2001 [cited 2025 Mar 3]. Available from: https://surface.syr.edu/cgi/viewcontent.cgi?article=1043&context=istpub.
32. Munot N, Govilkar SS. Comparative study of text summarization methods. Int J Comput Appl. 2014;102(12):33–37. doi:10.5120/17870-8810.
33. Liu Y. Fine-tune BERT for extractive summarization. arXiv:1903.10318. 2019.
34. Moratanch N, Chitrakala S. A survey on abstractive text summarization. In: 2016 International Conference on Circuit, Power and Computing Technologies (ICCPCT); 2016 Mar 18–19; Nagercoil, India: IEEE; 2016. p. 1–7. doi:10.1109/ICCPCT.2016.7530193.
35. Bird S, Klein E, Loper E. Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media [Internet]. 2009 [cited 2025 Mar 3]. Available from: https://www.nltk.org/book/.
36. Subramanyam Kalyan K, Rajasekharan A, Sangeetha S. AMMUS: a survey of transformer-based pretrained models in natural language processing [Internet]. 2001 [cited 2025 Mar 3]. Available from: https://mr-nlp.github.io.
37. Dai Z, Yang Z, Yang Y, Carbonell J, Le QV, Salakhutdinov R. Transformer-XL: attentive language models beyond a fixed-length context. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019; Florence, Italy. p. 2978–88.
38. Raffel C, Shazeer N, Roberts A, Lee K, Narang S, Matena M, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. J Mach Learn Res. 2020;21(140):1–67.
39. aphasie suisse | Home. Retrieved January 31, 2023. [Internet]. [cited 2025 Mar 3]. Available from: https://aphasie.org/.
40. Lin CY, Och FJ. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics—ACL '04; 2004 Jul 21–26; Barcelona, Spain; 2004. p. 605–12. doi:10.3115/1218955.1219032.

41. Pitler E, Nenkova A. Revisiting readability: a unified framework for predicting text quality. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing—EMNLP'08; Kerrville, TX, USA: The Association for Computational Linguistics; 2008 Oct 25–27. p. 186–95. doi:10.3115/1613715.1613742.
42. OpenAI API. Retrieved 2023 Jan 20. [Internet]. [cited 2025 Mar 3]. Available from: https://openai.com/api/.
43. Hugging Face—The AI community building the future. Retrieved January 20, 2023. [Internet]. [cited 2025 Mar 3]. Available from: https://huggingface.co/.
44. Cachola I, Lo K, Cohan A, Weld DS. TLDR: extreme summarization of scientific documents [Internet]. 2020 [cited 2025 Mar 3]. Available from: https://github.com/allenai/scitldr.
45. Schmitter P, Kirecci I, Gatziu Grivas S, Hanne T, Beck C. Transformation Compass für nicht-medizinische Supportprozesse (DE) [Internet]. 2003 [cited 2025 Mar 3]. Available from: https://digitalcollection.zhaw.ch/items/4cf03e1e-6ad0-45b4-8477-86fdf681a416.