

Prompt repairs and prompt patterns: Improving
prompt engineering for automated medical
reporting

Master Thesis Business Informatics

Submitted by Xinyu Mao 5194385

First Supervisor: Prof. dr. Sjaak Brinkkemper

Second Supervisor: Dr. Iris Beerepoot

February 12, 2024

Contents

1	Introduction	4
1.1	Gap in Knowledge	4
1.2	Problem Statement	5
1.3	Expected Contributions	6
2	Research Method	7
2.1	Research Questions	7
2.2	Design Science	8
2.2.1	Problem Investigation	9
2.2.2	Treatment Design	12
2.2.3	Treatment Validation	13
2.3	Threats to validity	14
2.3.1	Conclusion Validity	14
2.3.2	Construct Validity	14
2.3.3	Internal Validity	14
2.3.4	External Validity	15
3	Literature Review	16
3.1	Automated Medical Reporting	16
3.2	LLM & GPT in general prompt engineering	17
3.3	Prompt-based fine-tuning in healthcare domain	20
4	Method for SOAP report generation	25
4.1	Sample	25
4.2	Object	25
4.3	Method for generating and improving SOAP report by GPT	26
4.4	Pilot Trial Implementation	28
4.4.1	Analysis and Improvement of the Report	29
5	Repair Prompts and Prompt patterns for Automated Medical Reporting	32
5.1	Identifying type of issue	32
5.2	Repair prompts	32
5.2.1	RP-1 Brevity usage	35
5.2.2	RP-2 Short sentence usage	35
5.2.3	RP-3 Statement usage	36
5.2.4	RP-4 Abbreviation usage	36
5.2.5	RP-5 Hallucination prevention	37
5.2.6	RP-6 Timeline inclusion	37
5.2.7	RP-7 Medication inclusion	37
5.2.8	RP-8 Self-habit inclusion	38
5.2.9	RP-9 Surgery inclusion	38
5.2.10	RP-10 Clinical term usage	39
5.2.11	RP-11 Misplaced Subjective information correction	39

5.2.12	RP-12 Misplaced Objective information correction	39
5.2.13	RP-13 Misplaced Assessment information correction	40
5.2.14	RP-14 Misplaced Plan information correction	40
5.3	Prompt Pattern	41
5.3.1	Report Format Patterns	41
5.3.2	Report Content Adjustment	46
5.3.3	Report Checking	48
6	Evaluation	50
6.1	Method	50
6.2	Accuracy Results	54
6.3	Medical Summarizing Precision Score(MSPS)	56
6.4	Result of MSPS	58
7	Treatment Validation	61
7.1	Conclusion Validity and Internal Validity	61
7.2	Construct Validity	61
7.3	External Validity	61
8	Conclusion & Discussion	62
8.1	Findings	62
8.2	Limitation	62
8.3	Future work	63
	References	64
	Appendix A PDD Activity Table	69
	Appendix B PDD Concept Table	71
	Appendix C Example1 used in PP-1	73
	Appendix D Guideline1 in used in PP-2	74

1 Introduction

As is widely known, the primary duty of a doctor is to diagnose and treat patients. However, with the continuous improvement of the healthcare system, doctors nowadays spend a significant amount of time on tasks such as documenting patients' complaints, inputting Electronic Medical Records (EMRs), especially trauma surgeons (Golob Jr, Como, & Claridge, 2016). In the Netherlands, more than 90% of general practitioners commonly utilize electronic medical records (Jha, Doolan, Grandt, Scott, & Bates, 2008). In a survey of 2014 Massachusetts General Physicians Organization it was found that a quarter of medical staff working time is spent on administrative work on average (Rao et al., 2017). EMR enables the long-term storage of patients' medical histories. Even a doctor who is unfamiliar with a patient's past can quickly grasp their medical history and make a diagnosis through the use of electronic medical records. Additionally, EMRs can record patients' medication information. In the event of adverse drug reactions, doctors can promptly identify the source of the problem by referring to electronic medical records. Implementing an EMR system not only serves as a form of accountability for healthcare professionals but also has the potential to significantly improve the quality of healthcare services (Heart, Ben-Assuli, & Shabtai, 2017). Furthermore, the use of electronic medical records can to some extent facilitate communication between hospitals and patients.

However, the widespread adoption of electronic medical records has also increased the workload of doctors, resulting in longer working hours. What is worse, the administrative burden negatively impacts the satisfaction of their work, even patient-related duty (Rao et al., 2017). Research indicates that doctors who bear a significant administrative burden tend to have lower levels of professional satisfaction (Woolhandler & Himmelstein, 2014). To address the mentioned issue, some healthcare institutions have employed medical scribes to transcribe clinical encounter information into electronic medical records. While this can be an effective measure to alleviate the burden on doctors, it also introduces additional expenses. Considering the potential negative impact of increased administrative burden on doctors' job satisfaction and patient care, it is crucial to explore and implement more effective methods for documenting electronic medical records, thereby alleviating the burden on doctors and improving overall healthcare efficiency.

1.1 Gap in Knowledge

In the healthcare field, there has been increasing amount of research focused on formulating automated medical reporting from doctor-patient conversations to reduce recording time (Knoll et al., 2022). The research goal is to generate an EMR automatically based on the input of doctor-patient conversations audio. To archive this goal, the first step is to acquire the transcript from doctor-patient conversations. Afterwards, multiple recording methods such as

wearable devices, smartphones, integrated recording systems and digital voice recorder could be used to preserve the authentic doctor-patient conversations. Subsequently, employing speech recognition technology, the recorded audio files are transcribed into written text, thereby reducing the time physicians spend on manual typing. Implementing this approach necessitates rigorous evaluation and validation to ensure the accuracy and reliability of the transcriptions. Furthermore, privacy and confidentiality concerns should be addressed through appropriate measures, such as secure storage and data encryption, to safeguard patient information (Falcetta, de Almeida, Lemos, Goldim, & da Costa, 2023). The next step is to utilize model to generate EMR through the transcript. Considering the feature of using speech-to-text as an input item, it would be better to use Large Language Model (LLM) ased on Generative Pre-trained Transformer (GPT) to generate EMR. Figure 1 illustrates the process of automated medical reporting using GPT prompts. Firstly, the prompt instructor is trained using actual doctor-patient conversation transcripts and a standardized EMR. Next, the prompt pattern is extracted from the training data. Finally, the combination of GPT and the prompt pattern is utilized to generate the EMR from the transcript.

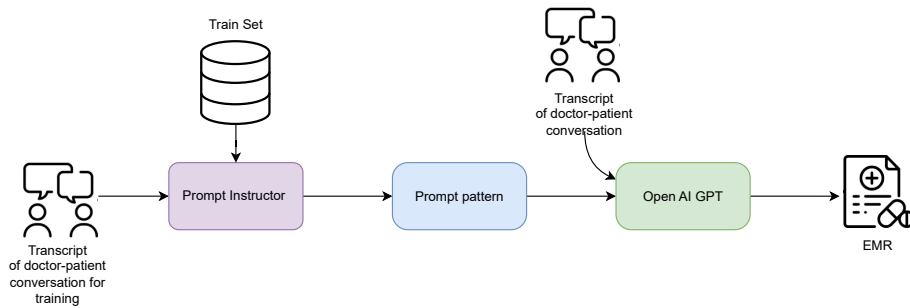


Figure 1: The process of automated medical reporting using GPT prompts.

1.2 Problem Statement

While the automatic formulation of doctor-patient conversations into EMR has garnered significant attention in the healthcare research domain, there is still a need for improvements in accuracy (Schukow et al., 2023). Extractive abstractive approach is one of the methods proposed by (Alsentzer & Kim, 2018) to formulate automated medical reporting. However, their research showed that the upper bounds on extractive summarization in medical text could not meet the requirement of a sufficient recall. The research on Automatic Dialogue Summary Generation, which is applied to the summary task based on dialogue, has been conducted (C. Liu, Wang, Xu, Li, & Ye, 2019). However, they did not explicitly model the properties of medical data. A recent study showed that the EMRs which are generated by the fine-tuning model are not preferred over the

EMRs manually created by humans (Giorgi et al., 2023). Therefore, there is a need for a more accurate method to automatically generate EMRs, replacing the current manual process. The Care2Report (C2R) research project, led by Professor. dr. Sjaak Brinkkemper from Utrecht University, endeavors to develop automated medical reporting techniques aimed at reducing the time required for documenting medical notes following patient consultations. By integrating advanced technologies such as natural language processing and speech recognition, the project aims to enable healthcare providers to generate comprehensive and accurate medical reports more efficiently. This thesis will use prompt engineering, focusing on the design of prompt patterns and fine-tuning of the GPT model, to enhance the accuracy of automatically generating EMR transcripts.

1.3 Expected Contributions

This thesis is expected to make contributions to both theoretical and practical aspects of the healthcare domain. On the theoretical front, it will focus on creating prompt patterns specifically tailored for the healthcare context. By delving into the intricacies of medical documentation, this research will shed light on effective strategies for generating prompts that align with the unique characteristics of healthcare reports. These theoretical contributions will enhance our understanding of prompt patterns and their role in improving the accuracy and comprehensiveness of automated reports.

On the practical side, this research will enhance the accuracy and quality of automated reports in healthcare by incorporating the created prompt patterns. By capturing the unique characteristics of medical documentation, these prompt patterns will improve the ability of automated systems to accurately interpret and capture medical information, resulting in more precise and reliable reports. This contribution will provide healthcare professionals with trustworthy automated reports, supporting better decision-making and patient care. Moreover, this research will reduce the time required for doctors to write reports. By leveraging effective prompt patterns and automated reporting techniques, the manual effort and time needed for report writing will be minimized. This reduction in workload will improve workflow efficiency, allowing doctors to allocate more time to direct patient care, research, and other essential tasks. The practical impact of this contribution lies in optimizing resource allocation and improving overall productivity and time management within healthcare settings.

In summary, this thesis will contribute to the theoretical understanding of prompt patterns in the healthcare domain and provide practical solutions to enhance the accuracy and quality of automated reports. By reducing the time burden on doctors for report writing, healthcare institutions can optimize their resources and improve the efficiency of medical documentation processes. These expected contributions have the potential to advance the field of automated reporting in healthcare, benefiting healthcare professionals and ultimately enhancing patient outcomes.

2 Research Method

2.1 Research Questions

In order to bridge the knowledge gap and achieve the research objectives, this paper has formulated the following main research question (MRQ):

- MRQ: How can experiments in prompt engineering for the improvement of automated medical reporting be executed?

To answer the main research question, a set of experiments will be executed including creating prompts pattern using GPT to improve the automated medical reporting. This research will include the experiment The main research question could be further decomposed into the following research questions (RQs).

- RQ1: What is the current landscape of application of LLM in the healthcare domain?

Having a good overview of the core research technique is the basis of the research. Thereby, the first step is to get knowledge of the related technology and how it has been used in healthcare domain. This sub-research question will be answered by related work.

- RQ2: What steps should be taken for improving automated medical reporting?

The second step is to create the framework of the whole process to enhance the competence of automated medical reporting. A clear understanding of macro processes is a prerequisite for improving automated medical reporting. A pilot experiment will be conducted. This sub-research question will be answered by creating a Process Data Diagram (PDD) of generating automated medical reporting.

- RQ3: What GPT prompt patterns could be formulated to improve for automated medical reporting?

After getting the knowledge of the background and overview of automated medical reporting, the next step is to develop a method to improve the automated medical reporting to increase the accuracy. This sub-research question will be answered by creating a set of prompt patterns of GPT-3. These prompt patterns are expected to fine-tuning EMRs in the perspective of completeness and accuracy. The related knowledge of prompt pattern will also be introduced in related work.

- RQ4: How to validate the prompt patterns in generating accurate electronic medical records?

Table 1: Method(s) used to answer RQs

	Literature Review	Design Science	Experiment
RQ1: What is the current landscape of application of LLM in health-care domain?	X	X	
RQ2: What steps should be taken for improving automated medical reporting?	X	X	X
RQ3: What GPT prompt patterns could be formulated to improve for automated medical reporting?		X	X
RQ4: How to validate the prompt patterns in generating accurate electronic medical records?		X	X

The last procedure of the research is to evaluate the prompt pattern in generating accurate EMRs. To answer this sub-research question, an experiment will be conducted. This experiment will use the transcript from the recording of district home care nurses-patients.

The whole process of the study can be completed using the design science framework. Table 1 shows the corresponding solutions for each problem.

2.2 Design Science

Design Science Research (DSR) is a research paradigm in which the engineering cycle as shown in Figure 2 is the framework used in information systems and software engineering design science, as proposed by Roel Wieringa. Design science is the design of artifacts in context (Wieringa, 2014), focusing on the development and validation of prescriptive knowledge in information science. Artifacts in context of design science could include models, methods, constructs, substantiations and design theories (March & Smith, 1995). The engineering cycle consists of three phases, namely problem investigation, treatment design and treatment validation. In the phase of problem investigation, it is crucial

to identify the stakeholders within the context and determine the objectives to be achieved. Once the problem has been clearly defined, the next step involves treatment design. This phase entails identifying and prioritizing the requirements that are most relevant to the stated objectives. Subsequently, the artifact is designed, integrating and addressing these identified requirements. The final step involves the validation of the designed artifact, providing valuable insights for future iterations and investigations. This iterative process, comprising investigation, design, and validation, constitutes the design cycle, which is embedded within the broader engineering cycle. The outcome of the design cycle is a validated treatment method that requires practical implementation and evaluation in real-world settings. These activities, based on the design cycle, collectively form the engineering cycle. This paper will adopt this research framework to guide a systematic and rigorous investigation. This research framework will answer RQ 1,2,3,and 4 in different sections.

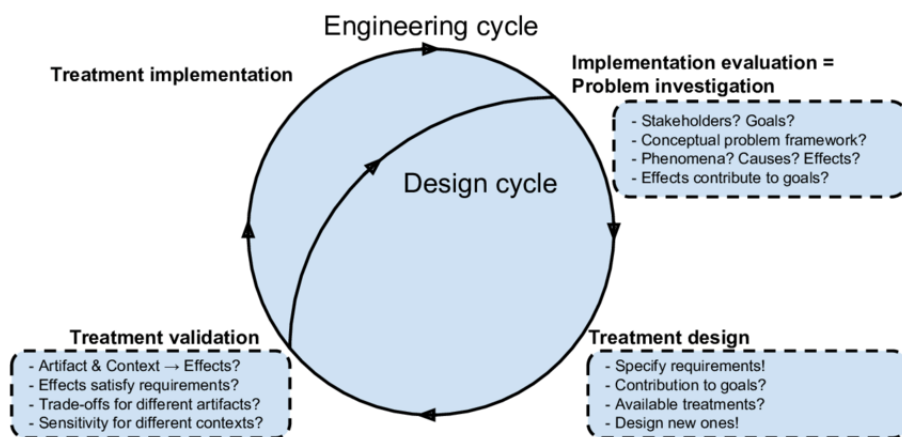


Figure 2: Wieringa-Engineering-cycle (Wieringa, 2014).

This study adopts the paradigm of DSR to accomplish the research objectives. Table 1 illustrates the relationship between Design Science as the main thread of this research and each research question along with their corresponding methods. In the context of this paper, the artifact within design science refers to the improving method of GPT-4, specifically within the context of the automated generation of EMR.

2.2.1 Problem Investigation

In the problem investigation phase of the design science approach, the primary objective is to gain a comprehensive understanding of the research project, encompassing its background and specific details. This phase involves conducting an extensive literature review to examine the current state of affairs and iden-

tify key challenges, which could answer RQ 1 and 2. The focus is on obtaining insights into the problem’s contextual framework, its stakeholders, and their respective requirements.

To address the aforementioned issues, a thorough literature review will be conducted. The first step is to acquire a comprehensive understanding of the background related to the automated generation of electronic medical records. The motivation behind this requirement is to alleviate administrative burdens on healthcare professionals, allowing them to allocate more time to their core expertise of diagnosing patients rather than engaging in administrative tasks. Therefore, the initial part of the literature review will revolve around exploring the administrative burdens in the healthcare domain. The primary focus will be to ascertain the presence of administrative burdens in this field and their impact on physicians, hospitals, and other stakeholders. Furthermore, an examination of the current research status regarding the utilization of GPT technology for generating electronic medical records will be conducted to determine the existing problems and areas for improvement. Lastly, a literature review will be conducted on the key technological area of prompt-based learning, which will contribute to the creation of high-quality prompt patterns. After confirming the content of the review, the practical process of literature review is shown in Figure 3. Besides academic literature, this paper also includes grey literature in the industry. Bridging the gap between academic research and professional practice is one of the key advantage to introduce grey literature (Elmore, 1991). Grey literature includes white book, website content, blog and it could be generated in the industry especially while applying the theory (Tom, Aurum, & Vidgen, 2013).

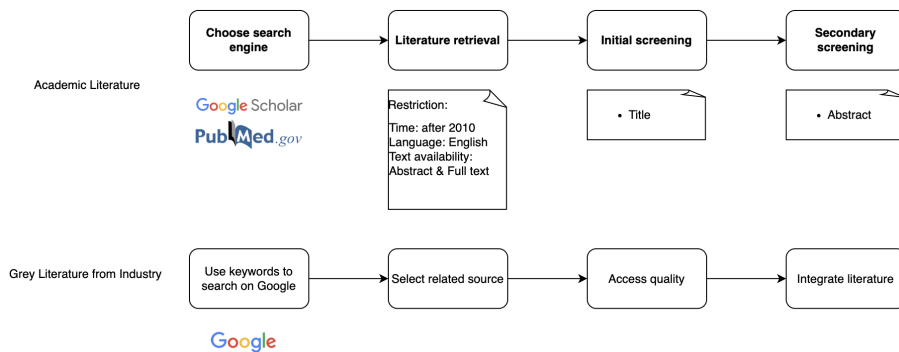


Figure 3: The process of Multivocal Literature review.

Academic literature review process

1. Choose search engine
All relevant literature sources will be searched from Google Scholar and

PubMed. Google Scholar is a free and easy tool for researchers to find scholarly information on any topic (Jacsó, 2005). Given this research is under the background of healthcare, Pubmed which is focusing on the MEDLINE database of references and abstracts on life sciences and biomedical topics is also included.

2. Literature retrieval

In this step, the search query will be determined based on the desired content for the literature review, and the following limitations will be imposed:

- Publications should have been published from 2010 onwards.
- Only articles written in English will be considered.
- Full text and abstract should be available for the articles, as this will facilitate the subsequent step of refining the literature.

By incorporating these restrictions, the literature search will be focused on recent and relevant publications in the English language, providing access to both the complete text and abstracts of the articles. This will lay the groundwork for refining the literature and selecting the most pertinent sources for the study.

3. Initial screening

The initial screening will contain reviewing article titles.

4. Secondary screening

Abstracts will be thoroughly examined to identify the required literature. Finally, the useful and pertinent literature will be determined for further analysis and inclusion in the study.

5. Analyze Literature

Analyzing literature involves synthesizing selected works within a specific topic. It's crucial to identify the methods used and the outcomes achieved in these works. By comparing these methods and results, one can determine which approaches contribute most effectively to understanding the holistic topic.

Grey literature review process

1. Keyword Search on Google

The first step is to determine the appropriate search keywords related to the research topic. For instance, keywords like "prompt pattern" and "automatic medical reporting generation" can be selected. These keywords are then used to perform a search on Google.

2. Source Selection

Upon executing the Google search, a list of potential sources is generated. It's important to note that Google search results may include advertisements, which could impact the relevancy of the results. Therefore, careful consideration is needed when selecting sources from the list.

3. Assess Source Quality

Once relevant sources have been identified, the next critical step is to assess the quality of these sources. A comprehensive checklist for assessing the quality of grey literature has been provided by (Garousi, Felderer, & Mäntylä, 2019). This checklist includes criteria such as the authority of the producer, methodology employed, objectivity, publication date, novelty, impact, and outlet type. Evaluating sources against these criteria ensures that only high-quality grey literature is included in the review.

4. Integrate Literature

After filtering and selecting valuable sources, the literature needs to be integrated into the review. This involves categorizing the selected sources based on their relevance to the research topic. It's essential to refer back to the research question and retain sources that are likely to provide meaningful insights.

2.2.2 Treatment Design

For treatment design, (Wieringa, 2014) aims to design new artifacts which could figure out the problem that investigated above. This phase will answer RQ 3 and 4. In the context of this research, a method which could formulate automated medical reporting should be accounted as the artifact. Based on the current achievement of Care2Report, basic pipeline of generating automated medical reporting has been verified. However, the accuracy of the EMR should be improved by creating prompt patterns used in GPT. Thus, this research will focus on the fine-tuning phase of automated medical reporting.

The design of suitable prompt patterns involves considering various factors, such as the desired level of specificity, the scope of medical conditions covered, the target audience, and the desired structure of the system's output. Striking a balance between providing sufficient context for accurate reporting and avoiding excessive prompting is crucial to ensure the system's flexibility and adaptability. In determining the desired level of specificity, the prompt patterns should be designed to capture relevant and precise information while avoiding overly generic or ambiguous prompts. The scope of medical conditions covered in the prompt patterns should align with the intended application of the automated medical reporting system, encompassing a comprehensive range of conditions while also considering any specific focus areas or specialties. Moreover, the design of prompt patterns should consider the desired structure of the system's output. This involves determining the format, organization, and level of detail in the generated reports. By designing prompt patterns that guide the system to produce output with a consistent and structured format, the usability and interpretability of the automated reports can be enhanced.

2.2.3 Treatment Validation

In the treatment validation phase, the objective is to evaluate the accuracy of the automated medical report generated by GPT using prompts. This validation involves both quantitative analysis of auto-assessment metrics and human evaluation. Quantitative analysis in this phase is conducted using three metrics: Precision, Recall and Medical Summarizing Precision Score(MSPS). Precision and recall are the usual metrics to measure the performance of the system. Precision assesses the system’s accuracy and relevance while minimizing false or extraneous details. The formula of Recall is shown in Equation 1. Recall measures the system’s ability to accurately capture and include all relevant information in the generated reports. The formula of Recall is shown in Equation 2. Since some words would have more than one synonym, using the metric of precision may underestimate the performance. MSPS is the evaluation method created by this research which will be elaborated in Section 6.

ROUGE and Bilingual Evaluation Understudy (BLEU) are the methods of automatic machine translation evaluation. ROUGE is a method used to automatically evaluate the quality of a summary (Lin, 2004). It compares the computer-generated summary to human-created ideal summaries and measures the accuracy of the summary. The formula of ROUGE is shown in Formula 3. BLEU computes a score ranging from 0 to 1, where a higher score indicates a better translation. BLEU would consider multiple reference translations (Papineni, Roukos, Ward, & Zhu, 2002). The formula of ROUGE is shown in Formula 4. However, in the real pilot trial, this ROUGE and BLEU would get low score due to the feature of GPT. On one hand, GPT might generate too much related but unnecessary information which may decrease ROUGE and BLEU. On another hand, it is difficult for GPT to generate the exactly same information as a doctor written. These two aspects will result in a very low score of ROUGE and BLEU. Thus, these two metrics are not applicable in this research.

Moreover, it is important to note that these automated assessment metrics are not perfect and may not capture all aspects of human judgment (Moramarco et al., 2022). Although the generated Electronic Medical Records (EMRs) will adhere to the standard Structured Operationalized Examination Protocol (SOEP), there is still room for sentence or phrase rephrasing to achieve variation. Therefore, human evaluation is also considered as part of the validation process. The three metrics mentioned above are all based on the human evaluation. Human should categorize the statement that generated by GPT corresponding to which is shown in human report. Afterwards, it is started to calculate the score of these three metrics.

The validation process involves a dataset of 7 records, with a specific allocation for validating and evaluating the system’s performance. Out of the 7 records, 5 will be dedicated to testing the accuracy of the GPT generated report. The other two records will be used as examples in the prompt. Thus, only

five GPT generated reports could be evaluated. The data that are evaluated come from the research of (van Zandvoort, Wiersema, Huibers, van Dulmen, & Brinkkemper, 2023).

$$Precision = \frac{TruePositives(TP)}{TruePositives(TP) + FalsePositives(FP)} \quad (1)$$

$$Recall = \frac{TruePositives(TP)}{TruePositives(TP) + FalseNegatives(FN)} \quad (2)$$

$$ROUGE-N = \frac{\sum_{gram \in cand} \min(count_{gram}(cand), \max(count_{gram}(ref)))}{\sum_{gram \in ref} count_{gram}(ref)} \quad (3)$$

$$BLEU = BP \times \exp\left(\frac{1}{N} \sum_{n=1}^N \log p_n\right) \quad (4)$$

2.3 Threats to validity

2.3.1 Conclusion Validity

- **Generalization:** The findings and conclusions of the research may not be generalizable to other contexts or populations beyond the specific data set used in the study.
- **Sample Size:** The sample size of 5 records may not be sufficiently large to draw definitive conclusions, and the results may vary with a larger or more diverse data set.

2.3.2 Construct Validity

- **Measurement Validity:** The measures used to assess recall and precision may not fully capture the intended constructs, leading to inaccuracies or limitations in evaluating the system’s performance. Besides, human evaluation could contain subjective factors.

2.3.3 Internal Validity

- **Confounding Factors:** Confounding factors such as the ability of the expression and the complexity of the case could influence the observed relationship between the prompt patterns and the accuracy of the automated

medical reporting system.

2.3.4 External Validity

- **Generalizability:** The findings of the research may not apply to other healthcare settings or populations due to differences in healthcare practices, data collection procedures, or technological infrastructure.
- **Time Sensitivity:** The effectiveness and performance of the prompt patterns in the automated medical reporting system may change over time as new technologies, practices, or guidelines emerge.

3 Literature Review

3.1 Automated Medical Reporting

In contemporary healthcare, the extensive utilization of information technology is unmistakable. EMR is a good example. EMRs are systematic electronic repositories of patient health data, encompassing medical histories, medication records, vital signs, laboratory results, and healthcare professionals' notes (Campanella et al., 2016). This section explores the role of EMRs in healthcare and its implications. Outpatient clinical documentation is the only task that is considered both valuable and onerous (Rao et al., 2017). In healthcare facilities, the application of information technology in healthcare has shown benefits in various areas such as automating medication, examination and ensuring standardised of the medical process. Accumulating evidence suggests that EMR could improve the quality of healthcare by improving time efficiency and adherence to guidelines, as well as reducing medication errors (Campanella et al., 2016).

However, the integration of information technology, notably EMRs, adds a substantial administrative workload for healthcare providers. Tasks such as prior authorizations, clinical documentation, and medication reconciliation have been particularly burdensome (Rao et al., 2017). Primary healthcare providers, in particular, invest a significant amount of time in these administrative tasks, leading to a considerable imbalance compared to specialists (Rao et al., 2017). For instance, primary care physicians in the United States spend nearly two hours on EMR-related tasks for every hour of direct patient care (Arndt et al., 2017).

To alleviate this administrative burden, medical scribes have been considered. These professionals record patient-physician encounters in real-time (Finley et al., 2018). Although medical scribes can enhance healthcare providers' productivity, they also incur substantial costs. For example, an academic family medicine practice in the U.S. hired two full-time medical scribes at a cost of \$79,500 annually (Earls et al., 2017). In response, the concept of an automated medical scribe has been proposed to offset these expenses.

In the healthcare domain, numerous endeavors have been made to develop automated medical reporting systems. An automated scribe that utilized various speech and language technologies has been introduced (Finley et al., 2018), although their system still had limitations in terms of functionality and scope. The process of forming an EMR is shown in Figure 4. The generation of EMRs through automation faces two primary challenges. Firstly, transcribing conversations between physicians and patients can be prone to errors, with the error rate increasing due to the specialized vocabulary within the medical domain. Recent research has developed a dialogue transcription system for physician-patient interactions, but it exhibited a high error rate of 18.3% (Chiu et al.,

2017). The second challenge lies in how to extract key information and convert it into structured EMRs once the correct transcription text is obtained. (Jiang et al., 2017) proposed that Artificial intelligence (AI) technologies and natural language processing (NLP) hold promise for enriching healthcare data. In this study, NLP leverages the textual version of physician-patient dialogues as input and aims to generate EMRs as output, training the model using rich real-world clinical data to enhance the accuracy of the generated EMRs.

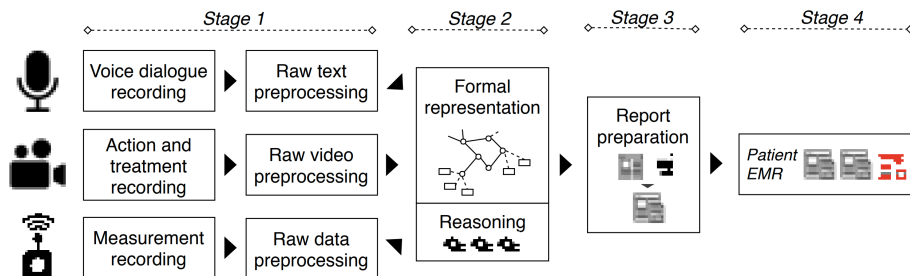


Figure 4: The process of automated medical reporting (Maas et al., 2021).

Care2Report is a recent software platform aimed at automating medical reporting (Maas et al., 2020). Figure 5 shows the dialogue summarization pipeline for reporting in healthcare. Firstly, utilizing audio recordings of conversations as input for speech transcription. This process involves converting spoken dialogue into written text using speech recognition technology. The transcribed text is then subjected to knowledge representation and ontology-based dialogue interpretation, enabling the extraction of meaningful information and context from the dialogue. Finally, the generated report is created and uploaded into the electronic medical record system, ensuring the availability of accurate and comprehensive documentation. The resulting EMR will adhere to the SOEP format employed by General Practitioners (GPs) in the Netherlands. These principles define four essential components for reporting consultations: Subjective (S), Objective (O), Evaluation (E) and Plan (P) (Cameron & Turtle-Song, 2002). Figure 6 illustrates the example of a paired doctor-patient conversation and clinical note in SOEP standard.

3.2 LLM & GPT in general prompt engineering

In recent years, large language model (LLM) has been enabled broadly in task performance (Srivastava et al., 2022). Generative Pre-trained Transformer (GPT) represents a remarkable breed of deep learning language models. Its defining characteristic lies in its pre-trained nature. Through extensive exposure to large amounts of textual data, GPT has absorbed the intricacies of language, contextual understanding and grammatical structure. This enables the model

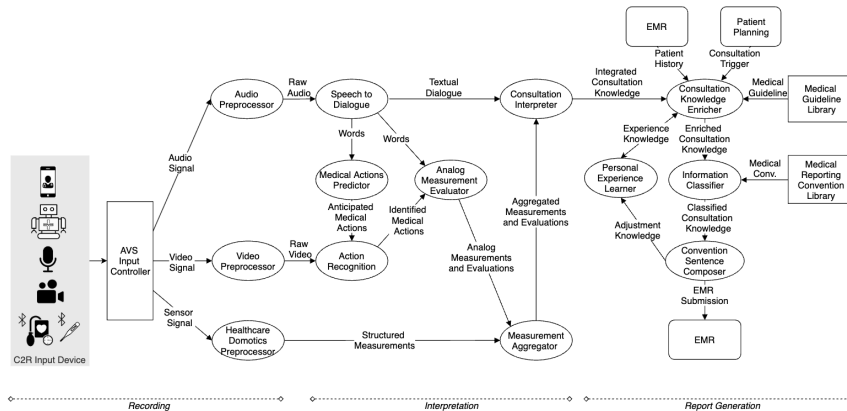


Figure 5: Functional architecture of the Care2Report system with components based on microservices (Maas et al., 2020).

to generate coherent, contextually consistent text based on input stimuli. GPT does not exist in isolation; it belongs to the broader category of LLM. LLM shares common architectures and traditions, but exhibit differences in size, capability, and scope. GPT occupies a unique position in the LLM ecosystem by virtue of its large number of pre-training programmes for different textual sources. OpenAI launched GPT-1 in June 2018 (P. J. Liu et al., 2018). Subsequent iterations benefited from training on more extensive language datasets, culminating in the latest version, GPT-4. GPT-4 has been fine-tuned with human intervention based on GPT-3.5, resulting in language output that closely approximates human speech (Wang et al., 2023). This milestone demonstrates the ongoing commitment to advancing generative pre-trained transformer models for natural language applications. GPT itself includes large amount of functions which do not need to learn during the training. In a recent BLEU scores report, it showed that GPT-3 has potential in baseline performance based on the experimental performance of variants GPT-3 using different prompt formats (Reynolds & McDonnell, 2021). Prompt engineering has come to the forefront since the introduction of GPT (Hu, Pan, Li, & Yang, 2023). This concept is aimed to optimize the language in a prompt in order to elicit the best possible performance (Zhou et al., 2022). It has the capability of providing a direct and friendly way for researchers to interface with LLMs. Accordingly, promoting engineering has been regarded as a usual method for doing NPL tasks (Schick & Schütze, 2020). Prompt engineering is the product of the development of LLM, providing more efficient LLM (Wang et al., 2023).

In general, prompt learning is a structured approach unfolds through five key steps, as depicted in Figure 8. Researchers commence this journey by selecting an appropriate pre-training model, thereby establishing the groundwork for

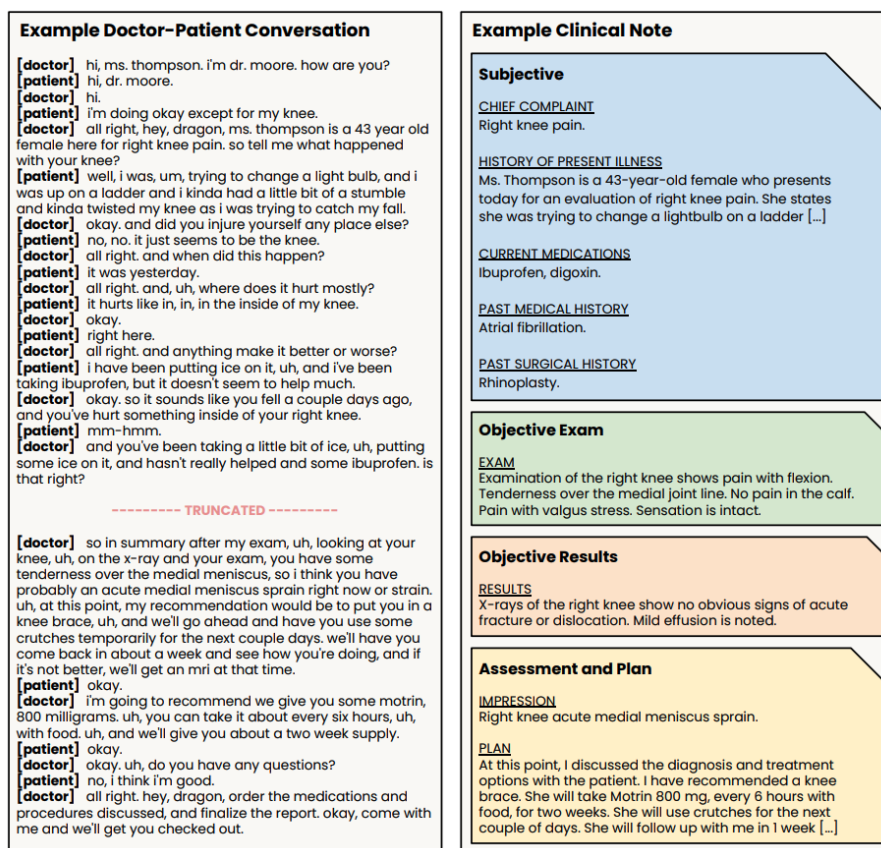


Figure 6: Example of a paired doctor-patient conversation and clinical note (Giorgi et al., 2023).

subsequent stages. Then comes the pivotal step of prompt engineering, where tailored prompts for downstream tasks are designed to align with the specific requisites of each task. Subsequently, in step three, task-specific responses are crafted, facilitating the model's ability to generate desired outputs. Step four delves into paradigm expansion, wherein methodologies are explored to enhance adaptability and yield improved results. Finally, in step five, the development of efficient training strategies empowers the model to learn effectively. At the heart of this approach, prompt engineering plays a vital role in ensuring the comprehensive and effective fulfillment of tasks. The design of the prompt significantly influences the entire model's performance. Prompt engineering is the approach to facilitate LLM through prompts (White et al., 2023). In the field of NLP, it usually contains two types. One is cloze prompts and another is prefix prompts shown in Figure 7. In the healthcare field, prefix prompts are frequently used. The following template serves as a simple example to demonstrate how a prompt

template can be structured. This template adheres to the standard EMR format in the Netherlands. When inputting patient records, this prompt template can be employed to construct the SOAP EMR. By utilizing this structured template as a prompt, it leads the generation of EMRs, ensuring that each section of the SOAP note is appropriately filled with related information.

```
1 Prompt Template:
2
3 "SOAP Template - {{Patient Name}} - {{Date}}
4
5 S (Subjective): {{Subjective Description}}
6
7 [Include patient's subjective complaints and symptom
8   description]
9
10 O (Objective): {{Objective Data}}
11
12 [Include objective observations, such as physical
13   examination and laboratory results]
14
15 A (Assessment): {{Assessment and Diagnosis}}
16
17 [Provide the doctor's assessment and diagnosis]
18
19 P (Plan): {{Treatment Plan}}
20
21 [Outline the treatment plan, medication recommendations,
22   and follow-up plan]"
```

In summary, prompt engineering represents a promising method within the domain of LLM. It places a strong emphasis on using prompts to guide the model, potentially becoming a crucial aspect of this field in the near future (Sorensen et al., 2022).

3.3 Prompt-based fine-tuning in healthcare domain

In the realm of natural language processing, the efficacy of pre-training has been prominently demonstrated, particularly in the domains of few-shot and zero-shot learning across a spectrum of general-domain tasks. This paradigm has gained substantial recognition and has been reinforced by recent research. Notably, a recent study harnessed a pre-trained language model, equipped with an impressive 1.37 billion parameters, to explore a direct yet highly effective approach for augmenting the zero-shot performance of pre-trained language models. The outcomes were remarkable, highlighting the proficiency of instruction-tuned models in excelling in previously unseen tasks within the zero-shot setting (Wei et al., 2021).

Fine-tuning, depicted in Figure 9-b, as another crucial phase, follows pre-

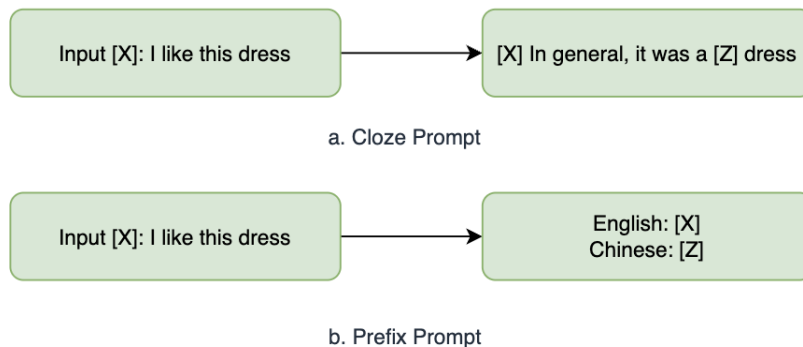


Figure 7: Process of prompt learning

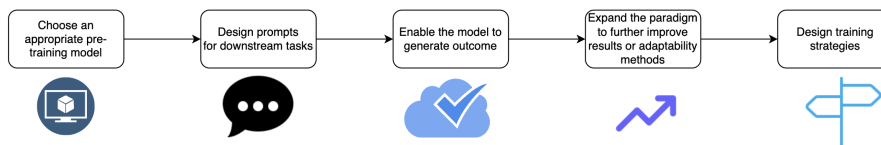


Figure 8: Process of prompt learning (Wang et al., 2023).

training and operates in a supervised manner, aligning the model with specific downstream tasks (P. Liu, Qiu, & Huang, 2016). In the context of automated medical reporting, this phase further refines the model, utilizing specialized medical reporting datasets to enhance its performance in generating tailored medical reports for clinical scenarios.

Prompt tuning, depicted in Figure 9-c, represents yet another significant phase in this structured approach. Prompt-tuning, inspired by the advancements in models like GPT-3, is aimed at aligning the training objective of downstream tasks with the pre-training stage, effectively converting them into a similar form (Feng et al., 2020). Prompt-based tuning customizes the model’s responses for different tasks through priming with natural language prompts—pieces of text combined with input to guide the language model in generating precise and context-aware medical reports. There are types of prompts, including manual prompts and automated prompts (P. Liu et al., 2023). Manual prompts, which are often crafted by human experts, serve as a means to guide models in generating task-specific results. This approach, involving human intervention, can be categorized into two distinct types: zero-shot prompts and few-shot prompts. Zero-shot prompts entail the direct utilization of manually designed prompts to generate results without prior testing of the model using samples. However, this approach has inherent limitations as it heavily relies on the performance of pre-trained models. In contrast, few-shot prompts involve preliminary

model testing using a limited number of samples before deploying prompts for content generation. The application of manual prompts has shown substantial enhancements, particularly in medical text enhancement tasks, when compared to other baseline models. On the other hand, automated prompts are generated using algorithms and techniques, such as discrete prompts and continuous prompts (Li & Liang, 2021). Discrete prompts operate within a discrete space, primarily involving the automatic search for templates within natural language phrases to guide the model’s output. Conversely, continuous prompts transcend text-based constraints and can incorporate spatial operations. They efficiently adjust parameters by modifying task-specific training data, making them a more efficient method for prompt-based fine-tuning. This multifaceted approach has demonstrated notable advantages, particularly within the realm of healthcare, where it offers logical and structured means of harnessing the potential of natural language processing models.

However, previous work has so far been unable to achieve competitive results on biomedical NLP tasks using domain-agnostic LLMs like GPT-3 (Gutiérrez et al., 2022). One of the reasons could be that the clinical texts are very different from those used in general NLP. Both grammatical and lexical aspects are very different from the texts used in general NLP (Wu et al., 2020). One solution could be scaling up language models, such as infusing medical knowledge. Recently, one study shows that conditioning on a few examples or a prompt could greatly improve task-agnostic, few-shot performance on various NLP tasks (Brown et al., 2020). Sarker, Qian, & Dong (2023) used data augmentation based on ChatGPT to enhance the accuracy of medicine identification and medication event classification (Sarker, Qian, & Dong, 2023). AugGPT is another model which reformulates each sentence in the training sample into multiple conceptually similar but semantically different samples for text data augmentation (Dai et al., 2023).

The recent advancements in the field of medicine have demonstrated the strong performance of GPT models when fine-tuned for medical applications. Recently, a medically-aware GPT-3 data labeler known as GPT-3-ENS has been introduced (Chintagunta, Katariya, Amatriain, & Kannan, 2021). This model effectively combines medical knowledge with an ensemble of GPT-3 models, primarily for the purpose of medical dialogue summarization. Agrawal, Heggelmann, Lang, Kim, & Sontag contributed to this progress by introducing guided prompt design techniques to guide LLMs towards producing structured outputs (Agrawal, Heggelmann, Lang, Kim, & Sontag, 2022). They also employed resolvers to map the LLM outputs to structured label spaces. Their tasks encompass a wider range of expected output structures, showcasing the versatility of GPT-3 in clinical NLP. Nori, King, McKinney, Carignan, & Horvitz reported remarkable results, indicating that with the development of GPT-4, it could achieve over 80% accuracy on the United States Medical Licensing Examination (USMLE) (Nori, King, McKinney, Carignan, & Horvitz, 2023). This performance surpasses that of GPT-3.5 and even specialized models like Med-

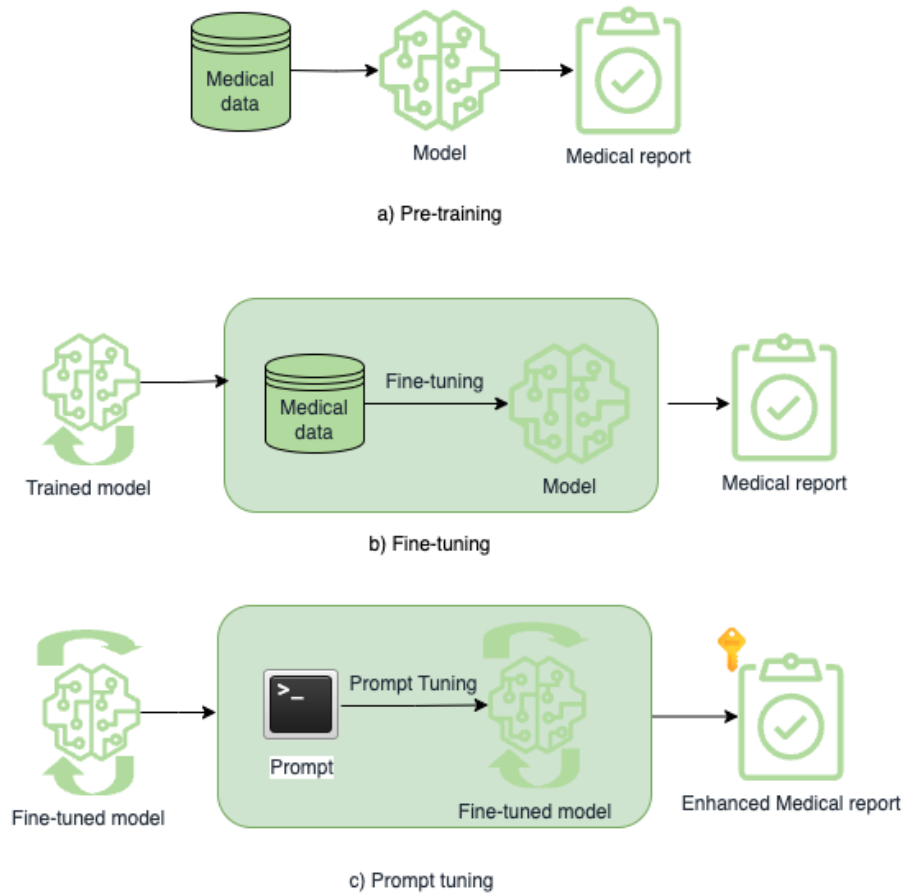


Figure 9: The process of pre-training, fine-tuning and prompt tuning on medical reporting task.

PaLM and Flanet, which are fine-tuned for medical knowledge.

Furthermore, (Sivarakumar & Wang, 2022) introduced a novel clinical NLP framework called HealthPrompt, which leverages prompt-based learning. They adapt task definitions using prompt templates, and this innovative approach yielded strong results in clinical text classification. In the context of privacy and security in medical data, (Z. Liu et al., 2023) presented DeID-GPT, a framework that utilizes GPT-4 to remove privacy information from medical data. Their framework outperformed other baseline experiments in safeguarding sensitive medical information. These recent studies collectively underscore the significant advancements in the field of medicine through the utilization of GPT models and innovative approaches, with the potential for even greater

achievements with the development of GPT-4.

4 Method for SOAP report generation

4.1 Sample

This paper will utilize transcriptions of patient-General Practitioner (GP) consultations dedicated to ear diseases in conducting the pilot trial (Maas et al., 2020). Furthermore, the GP will provide SOAP reports for each case. Both the transcripts and the SOAP reports are presented in Dutch.

4.2 Object

In this study, the object is the EMR generated by GPT. This paper employs the SOAP framework to formulate EMR, commonly utilized by many Dutch GPs to document EMRs. SOAP, an acronym for Subjective, Objective, Assessment, and Plan, serves as a structured format for organizing clinical information. The subjective component acts as a conduit for the patient's self-reported information, encompassing symptoms, the onset of the condition, pertinent medical history, medication use, and prior surgeries. This segment delves into the patient's emotional state, concerns, the intensity of their issues, and the impact on their relationships. Additionally, valuable insights from family members, case managers, and behavioral therapists may be incorporated (Cameron & Turtle-Song, 2002). The focus here is on the patient's perspective, and in some instances, medical staff may utilize first-person narration to convey information.

Conversely, the objective section pivots towards data derived from professional consultations and physical examinations, offering a complementary viewpoint to the patient's narrative. This section is grounded in factual information, presenting what the GP directly observes or witnesses. It encompasses quantifiable aspects such as visible signs, numerical measurements, sensory perceptions (smells or sounds), and any other empirical data (Cameron & Turtle-Song, 2002).

The "Assessment" section comprises the physician's diagnosis, which is based on a synthesis of the patient's subjective complaints and the objective findings obtained through medical examination.

Finally, the "Plan" outlines the proposed course of treatment and follow-up, addressing the steps to be taken in managing the patient's condition (Brenneman, 2001).

In general, there are some requirements when formulating SOAP. Figure 10 is one of the requirements from literatures. Furthermore, subjects, verbs, and function words are removed to resulting in shorter sentences (Smith, Megyesi, Velupillai, & Kvist, 2014).

Do	Avoid
Be brief and concise.	Avoid using names of other clients, family members, or others named by client.
Keep quotes to a minimum.	Avoid terms like <i>seems</i> , <i>appears</i> .
Use an active voice.	Avoid value-laden language, common labels, opinionated statements.
Use precise and descriptive terms.	Do not use terminology unless trained to do so.
Record immediately after each session.	Do not erase, obliterate, use correction fluid, or in any way attempt to obscure mistakes.
Start each new entry with date and time of session.	Do not leave blank spaces between entries.
Write legibly and neatly.	Do not try to squeeze additional commentary between lines or in margins.
Use proper spelling, grammar, and punctuation.	
Document all contacts or attempted contacts.	
Use only black ink if notes are handwritten.	
Sign-off using legal signature, plus your title.	

Figure 10: Guidelines for Subjective, Objective, Assessment, Plan (SOAP) Noting (Cameron & Turtle-Song, 2002).

4.3 Method for generating and improving SOAP report by GPT

Figure 11 shows the PDD generating and improving SOAP report by GPT. The whole process consists of three phases which are Data collection, Generating SOAP report and Improving SOAP report.

The initial phase is data collection, where a request for data download is submitted within the data management system of the collaborating medical institution. This step generates a TICKET, each marked with a unique identifier (Id), a Request detailing the nature of the data required, and a timestamp (time) indicating when the data request was made. Upon approval of this request, the relevant data can be downloaded from the system. The TRANSCRIPT SET comprises transcriptions of dialogues between doctors and patients, referred to as TRANSCRIPTS, as well as the SOAP reports written by physicians, hereinafter termed as HUMAN REPORT. Each transcription is correspondingly tagged with a No. To comply with the General Data Protection Regulation (GDPR), all transcripts have undergone a masking process to obfuscate personal information.

The subsequent phase entails the generation of a SOAP report. The process begins with arranging for a medical reporting pilot. Once the reporting platform is confirmed, the BASIC PROMPT which comes from the consolidation of literature describing the method of writing medical SOAP reports—is input into the GPT platform. Following the prompt, transcripts are entered into the system, which then produces a draft report, namely SOAP DRAFT. In accordance with the prompt, this SOAP report will contain four primary sections: SUBJECT, OBJECT, ASSESSMENT, and PLAN.

The last phase is improving the SOAP report. An initial task is to conduct a human evaluation, resulting in a document that encompasses the HUMAN REPORT, the GPT-generated report (DRAFT SOAP), and the evaluative comments from the human reviewer. Based on these assessments, a REPAIR PROMPT, including type and repair prompt, is created to enhance

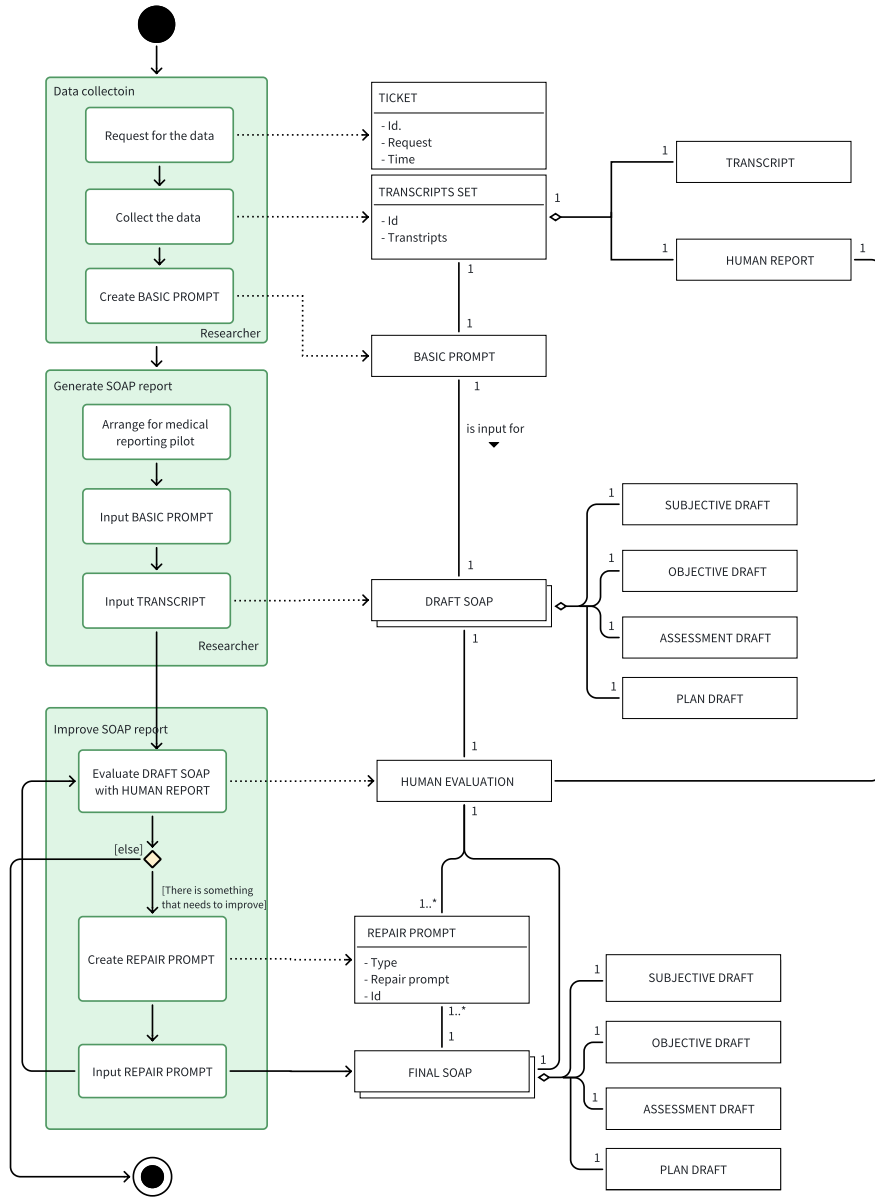


Figure 11: Method for improving EMR by GPT

the DRAFT SOAP. Each repair prompt is aimed at documenting a specific issue it addresses and its corresponding category. Inputting repair prompts into the system yields a revised report. At this juncture, it is necessary to deter-

mine whether the report meets the expected standards; if further refinement is needed, new repair prompts are created to guide the GPT toward an improved version. Once the report meets the expected outcomes, it is deemed the FINAL SOAP, thus concluding the pilot trial for a transcript.

4.4 Pilot Trial Implementation

This section delves into the implementation of a pilot trial based on the previously described Process-data diagram (PDD) model, using a doctor-patient transcript as an example. It illustrates the process of generating a report, creating prompts for improvement, and evaluating the final report. The initial step involves crafting a basic prompt, guided by existing literature guidelines. Below is the basic prompt created for this purpose.

```
1 Following are the transcripts in Dutch; the objective is
  to generate the SOAP report in English. In addition to
  the fundamental understanding of the SOAP format in
  medical reports, several general requirements must be
  adhered to:
2
3 Firstly, be succinct and precise, minimize quotations,
  use an active voice, and employ accurate and
  descriptive terminology. Begin each entry with the
  date and time of the session, documenting all
  interactions or attempted contacts.
4
5 Secondly, refrain from using names of other clients,
  family members, or individuals mentioned by the
  patient. Avoid terms like "seems" or "appears." Use "
  secondary" instead of "relevant." Avoid the term "
  suffer," as well as value-laden language, common
  labels, or opinionated statements. Employ medical
  terminology only if trained to do so.
6
7 Thirdly, if the exact nature of the problem is unclear,
  use diagnostic terms like "R/O" (rule out) or "EU" (
  etiology unknown), for example, "herpetic lesion EU; R
  /O contact phlebitis." The phrase "cannot be excluded"
  is also applicable, e.g., "migraine cannot exclude an
  aneurysm" or "cough cannot exclude pneumonia."
8
9 For the Subjective part, include symptoms, conditions,
  previous diagnoses, or brief statements describing the
  patient's current presentation. Refer to the
  following examples [example1],[example2].
```

4.4.1 Analysis and Improvement of the Report

The next step involves inputting two existing doctor-written reports as examples (example1 and example2). This serves as a learning reference for GPT to generate similar reports. The transcript is then input into GPT, and the following is the output from GPT.

GPT Report

- **S:** The patient has been asymptomatic after ear drops, now experiencing a return of initial symptoms with an itch. They clean their ears regularly, removing earwax with a cotton swab.
- **O:** AD (Right Ear): Skin flaking observed, no signs of infection evident. Tympanic membrane visibility is not mentioned.
- **A:** Diagnosed with bilateral otitis externa, the condition is recurrent and currently quiescent. No evidence of superimposed acute infection is noted at this time.
- **P:** Refer the patient to an otolaryngologist for specialized assessment. Discontinue current antibiotic treatment due to ineffectiveness and adverse reactions. Instruct the patient to keep the ear dry and clean, and to avoid inserting objects into the ear canal. Schedule a follow-up appointment to monitor the condition and discuss the otolaryngologist's assessment.

Human Report

- **S:** Since 3 days loop ear and pain again, now also on the right. Did see some improvement with antibiotics, but lots of side effects.
- **O:** ADS: otorrhoea and redness of ear canal
- **E:** Otitis externa both sides
- **P:** Iom pte referral ENT

The initial report lacked specific time markers, with the transcript indicating pain in the past three days, which was vague in the GPT-generated report. A repair prompt was created to address this timeline issue, leading to a more detailed account of the symptoms' onset and duration, although it still missed the recent intensification of pain. Another issue was the omission of medication details, which was rectified by including prompts about self-administration of medication, dosage, duration of usage, and its effects.

a. Repair Prompt for Subjective:

```
1 Review the transcript for any words related to
   date, time, or duration such as ""week, ""day,
   and include them in the 'Subjective' section".
```

The improved Subjective section: S: The patient reports that after using ear drops, they were initially without complaints. However, they are now experiencing a return of itching. The patient also regularly experiences earwax buildup and cleans it with a cotton swab. They mentioned that the condition has been ongoing for about four weeks, with the issue starting approximately six weeks ago.

Regarding the Objective section, GPT did not adequately reflect the doctor's observations. A targeted repair prompt was implemented to emphasize the need for details on the doctor's findings, leading to an updated report that included key information such as bilateral erythema and discharge. However, to ensure accuracy, another prompt was created to exclude patient-mentioned information from this section.

b. Repair Prompt for Objective:

1 Please do not include the information that the patient mentioned.

The improved Objective section: O: Examination of the ears revealed bilateral erythema and discharge, with the left ear more severely affected. Crystallization of exudate on the skin and severe scaling are noted. The patient has been using a cotton swab for earwax removal. No mention of fever or other systemic symptoms was made. Previous treatments with antibiotics were reported to be ineffective and poorly tolerated.

The Assessment section, while accurate, was overly verbose compared to the doctor's concise report. A prompt was created to refine the assessment, using a clinical term for a more succinct description.

c. Repair Prompt for Assessment:

1 For the assessment part, please use a clinical name to form the assessment.

The improved Assessment section: A: Bilateral Otitis Externa.

The initial Plan section included unnecessary explanations of the patient's condition, which were not relevant. A repair prompt was introduced to focus solely on the follow-up plan, aligning closely with the doctor's report in terms of content.

d. Repair Prompt for Plan:

1

Please do not include **any** other information that does not belong to the plan part. The plan is only focused on the follow-up plan, do not need to say again about the situation.

The improved Plan section: P: Schedule a referral to an otolaryngologist for further evaluation. Arrange a follow-up appointment post-specialist consultation to review treatment progress and next steps.

Final Report

- **Subjective:** The patient self-administered ear drops leading to a symptom-free period. However, itching has resumed. Regular earwax cleaning with a cotton swab is reported. Symptoms have persisted for about six weeks, with increased severity and new symptoms like pus and foul smell noted in the last four weeks. There was a temporary improvement with antibiotics, but significant adverse reactions occurred.
- **Objective:** Bilateral erythema and discharge in the ears were noted, more pronounced in the left ear.
- **Assessment:** Bilateral Otitis Externa.
- **Plan:** Schedule a referral to an otolaryngologist for further evaluation. Arrange a follow-up appointment post-specialist consultation to review treatment progress and next steps.

With this, the process of generating a SOAP report from a transcript in the pilot trial concludes, demonstrating the effectiveness of iterative prompts in refining the report to align closely with the medical professional's original document.

5 Repair Prompts and Prompt patterns for Automated Medical Reporting

5.1 Identifying type of issue

The previous section outlined the progression of our pilot test, where we identified several key problems in GPT's generation of accurate medical reports. During the pilot trial, we first summarized the issues discovered and the effective repair prompts used. Subsequently, we extracted prompt patterns from the useful repair prompts.

Firstly, different kinds of issues we spotted, demonstrating how these patterns were applied to enhance report accuracy, as shown in Figure 12.

The most common issue was GPT skipping over some details that the human report mentioned. This happened for two reasons: sometimes, the missing information wasn't in the transcripts—so, it makes sense GPT didn't include it. Other times, GPT just missed things that were actually said. We can fix this by giving GPT some extra hints, or "repair prompts" to improve the results.

Another problem was outright errors—when GPT gave us information that didn't match the human report or what was said in the transcripts. Take this case: a patient mentioned having ear pain for four months, which worsened in the last few days. GPT only noted the four-month duration, missing why the patient came in. Here too, repair prompts could guide GPT to focus on the recent worsening, which is really what matters.

Then there's the issue of adding extra stuff—details that weren't in the human report. Sometimes these were things found in the transcripts but weren't key points. Other times, GPT made up details without any source. Either way, we'll need repair prompts to help GPT stick to what's essential.

We also noticed GPT not using common abbreviations that doctors rely on to save time. Since reports are for doctors and patients alike, they should be easy to read. We're thinking of using a list of these abbreviations to teach GPT what to use.

Lastly, GPT's reports can be wordy, more so than what doctors write. We're looking at ways to get GPT to cut down on the word count without missing the important parts.

5.2 Repair prompts

The complete set of repair prompts utilized during the pilot trial to enhance the quality of the report generated by GPT are shown in Table 2. For conve-

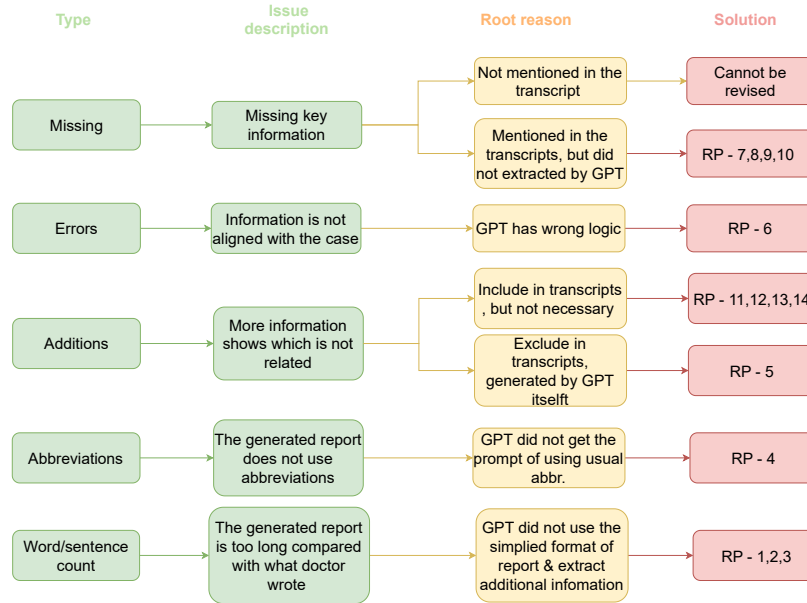


Figure 12: Type of issue.

nience, 'Repair Prompt' will be referred to as RP throughout this section. These RPs are grouped into three main categories: text related, medical related, and category related. The text related category addresses issues specific to textual content in report summarization, regardless of whether it pertains to medical summarization or not. The medical-related category is tailored specifically to issues unique to the medical field. Finally, the category-related group focuses on resolving issues such as misplaced information within different sections of the report. Each category is designed to address distinct aspects of report generation, ensuring clarity and accuracy in the documentation process.

Type	Name	Repair prompt	Description
Text related	RP-1: Brevity usage	Be brief and concise, keep quotes to a minimum, use an active voice, use precise and descriptive terms.	This is dedicated to formulating the simplified report. Once the sentence for the SOAP part is more than 5, this repair prompt should be executed again automatically.
	RP-2: Short sentence usage	Omit subjects, verbs, and function words resulting in shorter sentences.	This is used to make the report more simple and readable.
	RP-3: Statement usage	Using bullet points without subjective to formulate the report	This is one way to keep the report looks structured.
Medical related	RP-4: Abbreviation usage	If the report includes the following words, please use the abbreviation of them. PCM / Patient is PA or PT; Hospital is HSP; Right ear is AD; Left ear is AS	Medical staff usually use these abbreviations in the reports.
	RP-5: Hallucination prevention	Do not mention anything that is not shown or can not be diagnosed according to the transcripts.	Ensure that the report is in the minimal unit, ruling out anything that would obey the facts.
	RP-6: Timeline inclusion	Review the transcript for any words related to date, time, or duration such as "week", "day", and include them in the "Subjective" section.	GPT does not do well in identifying the timeline. Use this prompt to make sure that GPT will identify all the dates and times. In this way, helps GPT to generate the right subjective with the timeline.
	RP-7: Medication inclusion	Include details of patient self-medication events, specifying the medication dosage, duration of treatment, and the results observed.	The subjective part usually includes the medication situation. This prompt will help GPT to identify when and how much and what medicine the patient took before and had what result.
	RP-8: Self-habit inclusion	Include any of the patient's habits that relate to their symptoms in the subjective part of the report.	Subjective part usually includes patient self-habit. This prompt helps GPT include information about the patient self-habit in the subjective part.
	RP-9: Surgery inclusion	Please include the surgery history details about the patient, if any	Subjective part usually includes the patient's surgery situation. This prompt helps GPT include information about the patient's surgery situation in the subjective part.
Category related	RP-10: Clinical term usage	Employ clinical term to articulate the assessment. Repeat this prompt if the assessment exceeds one sentence.	Once the assessment part includes more than one sentence, this prompt should be executed automatically.
	RP-11: Misplaced Subjective information correction	Only incorporate details explicitly mentioned in the transcript to maintain relevance and accuracy.	Make sure Subjective only mentions the relative information.
Plan information correction	RP-12: Misplaced Objective information correction	Restrict the objective section to include only information that is directly observable and omit any plans or subjective information from patient.	Avoid mentioning any plans in the assessment section to ensure it contains only diagnostic conclusions and relevant findings.
	RP-13: Misplaced Assessment information correction	Avoid mentioning any plans in the assessment section to ensure it contains only diagnostic conclusions and relevant findings.	Make sure Assessment only mentions the relative information.
	RP-14: Misplaced Plan information correction	Ensure the plan section exclusively outlines the future management strategies without reiterating assessment or subjective/objective data.	Make sure Plan only mentions the relative information.

Table 2: Repair prompts for automated medical reporting

5.2.1 RP-1 Brevity usage

This is dedicated to formulating the simplified report. Once the sentence for the SOAP part is more than 5, this repair prompt should be executed again automatically. The following is the repair prompt.

1 Be brief and concise , keep quotes to a minimum ,
use an active voice , use precise and
descriptive terms .

Report Before RP-1

S: Patient reports chronic issues with the left ear, leading to significant discomfort and discharge, with recent worsening of symptoms. Complaints of pus and a foul odor, as well as nausea from the odor. The patient has had these symptoms for around six weeks and has a history of ineffective treatment with antibiotics.

O: Examination reveals redness and discharge in both ears, more severe in the left ear. The discharge has led to flaking and crusting around the ear. The patient has previously tried antibiotics and topical treatments with no significant improvement.

A: Possible severe otitis externa or complicated middle ear infection. The patient's condition has not improved with standard treatments, indicating the need for specialized ENT evaluation.

P: Refer the patient to an otolaryngologist for further assessment and treatment. Discuss with the patient the referral process and potential wait times. Provide interim care with a fresh dressing for the affected ear and instructions to return if the condition worsens before the specialist appointment.

Report After RP-1

S: Patient presents with a 6-week history of left ear discharge and pain, noting a foul odor and morning pus.

O: Left ear exhibits redness and discharge with crusting; right ear is red but less affected.

A: Chronic otitis with recent exacerbation, unresponsive to antibiotics.

P: Referral to otolaryngology for specialized assessment. Applied new dressing and instructed on care until specialist visit.

It is obviously that a comparative analysis of reports pre- and post-revision clearly demonstrates a reduction in length and increased succinctness after the implementation of the repair prompt.

5.2.2 RP-2 Short sentence usage

This is used to improve the wording of the report. The following is the repair prompt.

1 Omit subjects , verbs , and functional words resulting
in shorter sentences .

Report Before RP-2 Reported recurring ear condition last Saturday. No recent medical attention. Expressed dissatisfaction with previous treatments, **including** Sofradex ear drops. Mentioned self-administering [name of medication] four occasions past six months **with** favorable results, but the current episode does not respond to medication. Additionally, noted sore throat, experienced previously, suggesting a chronic issue.

Report After RP-2 Reported recurring ear condition last Saturday. No recent medical attention. Expressed dissatisfaction with previous treatments, Sofradex ear drops. Mentioned self-administering [name of medication] on four occasions past six months, with favorable results, but the current episode does not respond to medication. Additionally, noted sore throat, experienced previously, suggesting a chronic issue.

Upon comparison, it is easy to see that the functional words which were bolded in the paragraphs before the prompt repair have been removed after the execution of the repair prompts. This has made the report more concise.

5.2.3 RP-3 Statement usage

This category is aimed at distilling the report to its essence. By employing bullet points and avoiding subjective language, the report's structure is maintained with clear, concise statements. The specific repair prompt for simplification is as follows:

1 Use bullet points without subjective language to
formulate the report .

This method ensures that the report remains well-organized and straightforward.

5.2.4 RP-4 Abbreviation usage

This section addresses the standardization of terms within the report. When the report includes certain terms, their standardized abbreviations should be used to maintain consistency. The specific abbreviations repair prompt is as follows:

1 PCM or Patient is abbreviated as PA or PT;
2 Hospital is abbreviated as HSP;
3 Right ear is abbreviated as AD;

4 | Left ear is abbreviated as AS.

Medical staff are advised to consistently use these abbreviations in reports to ensure clarity and uniformity.

5.2.5 RP-5 Hallucination prevention

This category pertains to the content that should be included in the report. The directive is to omit any information that cannot be observed or diagnosed from the transcripts. Here is the repair prompt:

1 | Do not mention anything that is not shown or cannot
2 | be diagnosed according to the transcripts.
3 | Ensure that the report is presented in the minimal
4 | unit, excluding **any** details that would contradict
5 | the facts.

This ensures that the report remains factual and concise, avoiding the inclusion of unnecessary or unsubstantiated information.

Following are the repair prompts specific to the Subjective section, to be utilized when encountering particular issues.

5.2.6 RP-6 Timeline inclusion

GPT occasionally encounters difficulties in recognizing temporal expressions and establishing the chronology of events. To address this, a review of the transcript for any references to dates, times, or duration, for instance, 'week' or 'day,' is recommended within the 'Subjective' section. This process facilitates GPT's ability to accurately map temporal words to their respective events. The specific repair prompt to guide this enhancement is as follows:

1 | Review the transcript **for any** words related to **date**,
2 | time, or duration such as ""week, ""day, and
3 | include them in the "Subjective" section.

This assists GPT in generating a report with an accurate and detailed timeline.

5.2.7 RP-7 Medication inclusion

In the initial configuration of GPT, the model does not inherently recognize the necessity of incorporating medication details into the subjective portion of reports, a practice commonly observed in human-generated reports. It is essential to meticulously record any instances of medication self-administration by

patients, encompassing the dosage, duration, and resultant effects. The designated repair prompt to address this oversight is delineated as follows:

1 Include details of patient self-medication events ,
 specifying the medication dosage , duration of
 treatment , and the results observed .

This prompt ensures that GPT accurately reflects medication situations, thereby enhancing the subjective component of the report with essential details on medication intake and its effects.

5.2.8 RP-8 Self-habit inclusion

In the initial GPT model configuration, there is an omission of self-habit information that is typically documented in human-generated reports. Therefore, it is imperative to include details of the patient's personal habits when they bear relevance to the symptoms presented. The repair prompt designated for this aspect of the report is outlined as follows:

1 Include **any** of the patient's habits that relate to
 their symptoms in the subjective part of the
 report .

This addition helps GPT to enrich the subjective component of the report with pertinent personal habits.

5.2.9 RP-9 Surgery inclusion

: In the initial iteration, the GPT model does not account for the inclusion of surgical information that is commonly recorded in human-authored reports. Consequently, recording the patient's surgical history is of critical importance. The corrective repair prompt for integrating this section is as follows:

1 Please include the surgery history details about the
 patient , **if any**

This guideline assists in ensuring the subjective part of the report reflects all surgical contexts accurately.

The following outlines the repair prompts to be applied within the objective section when specific issues arise.

5.2.10 RP-10 Clinical term usage

In the initial model of GPT, the assessment section is generated as a narrative passage. Within a medical context, it is paramount that the assessment is conveyed using appropriate clinical terminology. The repair prompt to guide this requirement is as follows:

1 `Use clinical term or statement to formulate the
assessment.`

This ensures that the assessment part of the report is precise, professional, and succinct.

5.2.11 RP-11 Misplaced Subjective information correction

This category of repair is directed at ensuring that information not pertinent to the subjective section is not included therein. For example, details that are inherently objective should be confined to the objective section of the report; their inclusion in the subjective section would be inappropriate. This distinction is separate from previously mentioned additional information, which refers to content not found in the transcripts. The present category concerns the accurate allocation of relevant content to the appropriate sections of the report. The repair prompt to correct such misplacements is specified as follows:

1 `Only incorporate details explicitly mentioned in the
transcript to maintain relevance and accuracy.`

This ensures that the report is focused and avoids unnecessary speculation or assumptions.

5.2.12 RP-12 Misplaced Objective information correction

This repair prompt is typically employed to address instances where the objective section erroneously includes prospective plans or data not directly observed from the patient. The appropriate corrective measure for this section is outlined as follows:

1 `Restrict the objective section to include only
information that is directly observable and omit
any plans or subjective information from patient.`

This tip aligns with medical report guidelines to keep the objective part strictly evidence-based.

5.2.13 RP-13 Misplaced Assessment information correction

Initially, GPT lacks clarity regarding the specific content that should be encompassed in the assessment section. It is therefore necessary to instruct GPT that when composing the assessment portion of the report, it should not incorporate any information pertaining to future plans. The designated repair prompt for guiding this section is as follows:

1

Avoid mentioning **any** plans in the assessment section to ensure it contains only diagnostic conclusions and relevant findings.

This directive ensures that the assessment remains focused solely on the current clinical evaluation.

The following outlines the repair prompts to be applied within the plan part when specific issues arise.

5.2.14 RP-14 Misplaced Plan information correction

This repair prompt addresses the issue where the plan section of the report erroneously includes details from the assessment or subjective/objective observations. The corrective repair prompt for this section is as follows:

1

Ensure the plan section exclusively outlines the future management strategies without reiterating assessment or subjective/objective data.

This maintains the distinct purpose of the plan section, which is to propose subsequent steps for patient care.

5.3 Prompt Pattern

In the evolving world of artificial intelligence interfaces, improving the way we create prompts is essential for better interaction with users in various fields. The important research by (White et al., 2023) has introduced a set of prompt design patterns specifically for software engineering tasks. This study aims to apply the structure of these prompt patterns, as outlined by (White et al., 2023), to the medical area. A complete prompt pattern in this context includes several key parts, which are based on the structure shown in the provided image:

- A clear **name and category** that uniquely identifies the prompt pattern, making it easier to talk about and reference in the context of medical informatics.
- The **intent and context** section that explains the specific challenges the pattern is meant to tackle, along with its goals.
- A section on the **motivation** behind the pattern, which clarifies why it's important and what problem it helps to solve.
- A detailed description of the **structure and key ideas**, covering the essential information needed to guide the language model to the desired outcome.
- Examples of **example implementation**, showing how the pattern can be put into practice.
- An evaluation of the **Consequences**, discussing the advantages and any potential downsides of using the pattern, as well as advice on how to modify it for different situations in the medical field.

Prompt patterns for medical use, as identified during pilot trial, are systematically broken down in Table 3. These patterns were developed from a collection of prompts that have proven effective in generating medical reports. The process involved analyzing various prompts and selecting those that consistently enhanced the quality of the generated reports. For ease of reference, these prompt patterns will henceforth be abbreviated as PP. The goal of cataloging these patterns is to provide useful guidance for future research, not just for crafting medical reports but for broader applications in healthcare as well. This research also included [X], [Y], etc template placeholders for the pattern implementation inspired by (Sorensen et al., 2022).

5.3.1 Report Format Patterns

The following section delves into a detailed discussion of each prompt pattern, based on the previously provided structure. The first category to address is report format, which primarily concentrates on prompt patterns that bring about

Category	Name
Report format	PP-1: Example reference PP-2: Guideline principle PP-3: Context constraints PP-4: Language style
Report content adjustment	PP-5: Unnecessary statement exclusion PP-6: Necessary statement inclusion PP-7: Statement correction
Report checking	PP-8: Requirement validation

Table 3: Classification for prompt patterns in medical context

changes to the overall structure of a report, rather than being limited to modifications of just one part.

- **PP-1 Example reference**

- **Intent and Context:** The SOAP report is used as a unique identifier to establish a standardized format for medical documentation.
- **Motivation:** The standardized format ensures consistency and supports effective patient care management.
- **Structure and Key Ideas:**

1 Please use **format** in [X] as example(s)

- **Example Implementation:**

1 Please use **format** in [example1] as
example(s)
2 [X]: [example1]

The example1 in the above is shown in Appendix C.

Initial report

Subjective: The patient reports that they have had a discharge from the ear for approximately three to four weeks, with increasing pain and issues in both ears. The patient has been experiencing pus and a foul odor from the ear, particularly in the mornings, causing nausea. There has been some temporary improvement with antibiotics, but with significant side effects. The patient also mentions having had an adverse reaction to antibiotics in the past, with blisters from medication taken for a foot condition. The patient expresses a strong preference not to be referred to a specific hospital due to negative

past experiences and wishes to be referred to another hospital² if possible.

Objective: Upon examination, the care provider notes redness in the ear canal and discharge in both ears. The left ear appears to be worse, with yellow discharge noted. The care provider observes that the discharge has been crusty up to the hairline and that the substance from the ear crystallizes on the skin.

Assessment: Bilateral otitis externa is indicated by the symptoms of otorrhea (ear discharge) and redness in the ear canal.

Plan: In consultation with the patient, a referral to an ENT (KNO in Dutch) specialist is planned. The patient is instructed to call a specific number for the referral and is advised to follow up if the wait time for an appointment seems prolonged. In the meantime, the patient is provided with gauze and a bandage to manage the discharge.

Final report after using PP-1

S (Subjective): The patient has been experiencing ear discharge and pain for about four weeks, worsening over the past three to four days. Initially, there was some improvement with antibiotics, but the patient suffered significant side effects and now expresses a reluctance to continue with antibiotics. The discharge is described as pus-like and foul-smelling, particularly in the morning, which induces nausea.

O (Objective): Examination shows redness and discharge in both ear canals, with the left ear being more severely affected. The discharge is noted to crystallize on contact with the skin and has formed crusts up to the hairline.

A (Assessment): Bilateral otitis externa is suggested by the symptoms of otorrhea and redness in the ear canal.

P (Plan): The care provider has decided to refer the patient to an ENT specialist for further assessment and treatment. The patient is instructed to contact the referral number and is provided with gauze and a bandage to manage the discharge. It was advised to follow up if the wait time for the specialist is prolonged. The patient's preference for hospital referral was discussed and noted.

- **Consequences:** Take Subjective for an example, after the input of the example, GPT got to know that the discussion of follow-up plan should not be included in subjective part. What is more, the whole report is shorter than before. Thus, an example prompt pattern is vital to GPT model in order to generate the better report.

PP-2 Guideline principle

- **Intent and Context:** The report should adhere to clinical documentation guidelines and regulatory standards.
- **Motivation:** The report is designed to avoid the issues already men-

tioned in the guidelines, with the aim of making it as professional as possible based on the available information.

– **Structure and Key Ideas:**

```
1 Please make sure that the [SOAP] that you  
generated meets every requirement on [X]
```

– **Example Implementation:**

```
1 Please make sure that the SOAP that you generated  
meets every requirement on [guideline1]  
2 [X]: [guideline1]
```

Initial Objective Upon visual inspection, the clinician observes some residual debris in the ear, indicative of past inflammation but notes no current signs of infection. The patient completed a longer-than-usual course of antibiotic ear drops, taking them for about ten days without complications.

Final Objective after using PP-2 Examination shows no active inflammation. Residual debris from a previous condition is noted in the ear canal. Patient completed a 10-day regimen of otic antibiotics.

The guideline1 used in the above is shown in Appendix D.

- **Consequences:** This alignment prevents the basic errors mentioned in the guideline and enhances the accuracy of medical reports. Consider the objective section as an example; its final version is significantly more concise compared to the initial one. Conciseness is also a key requirement in the guideline. Therefore, providing GPT with a guideline is crucial when it is tasked with formulating content. With specific guidelines, GPT can tailor the results to meet the outlined requirements more effectively.

PP-3 Context constraints

- **Intent and Context:** SOAP report has specific field which would be better to constrain the context. In (van Zandvoort et al., 2023) research, they also use context as prompt to make more accurate report.
- **Motivation:** Medical terms could be generated under the context of medical reports.

– **Structure and Key Ideas:**

```
1 Please make sure that the [SOAP] that you
   generated is in the context of [X]
```

– **Example Implementation:**

```
1 Please make sure that the SOAP that you generated
   is in the context of formulating a medical
   report
2 [X]: formulating a medical report
```

– **Consequences:** Specificity in reporting prevents generalizations and improves the relevance of the care plan.

PP-4 Language style

– **Intent and Context:** Merely relying on context and examples results in improvements at a general level for the report. For more specific linguistic handling, detailed instructions are necessary.

– **Motivation:** The inclusion of medical language instruction will make the report more closely resemble a human report.

– **Structure and Key Ideas:**

```
1 Please use [X] to formulate [SOAP] to avoid [Y]
```

– **Example Implementation:**

```
1 Please use statement, active voice to formulate
   SOAP to avoid long sentences
2 [X]: statement
3 [Y]: long sentences
```

Subjective after using PP-4 Patient experiences ear discharge and pain for four weeks; worsening in the last three to four days. Initial improvement was noted with antibiotics, but severe side effects followed. Patient is now reluctant to continue antibiotics. Reports pus-like, foul-smelling discharge in mornings, causing nausea.

– **Consequences:** This example, previously used in PP-1, has been simplified into concise statements rather than lengthy sentences. As a result, the subjective part of the report has become more readable with the application of this prompt pattern. The report will be generated in a statement style rather than lengthy descriptive text.

5.3.2 Report Content Adjustment

This section primarily focuses on adjusting the content of the report. It involves refining the details through various prompts, including adding missing information, deleting superfluous content, and correcting inaccuracies.

- **PP-5 Unnecessary statement exclusion**

- **Intent and Context:** Including only essential and correct information avoids inconvenience for both patients and doctors.
- **Motivation:** The goal is to concisely convey the complete situation using the minimum necessary information.
- **Structure and Key Ideas:**

```
1 Please remove [X] that [Y]
2 [X]: the information you want to remove
3 [Y]: the explanation of the information why they
   need to be removed
```

- **Example Implementation:**

```
1 Please remove the information that does not show
   or can not be inferred from the transcripts
2 [X]: the information
3 [Y]: does not show or can not be inferred from the
   transcripts
```

Initial Plan Advise to extend the use of otic drops for one week to promote canal cleanliness. Patient to monitor for symptom recurrence and initiate prompt treatment if similar sensations arise. Recommended utilization of online services for appointment scheduling and prescription refills. Advised against at-home cerumen removal and to seek professional otic irrigation if necessary.

Final Plan after using PP-5 Continue otic drops for one additional week to aid in the ear's self-cleaning process. The patient is advised to refrain from inserting any objects into the ear for cleaning purposes and to consult a healthcare professional if experiencing wax buildup or a recurrence of symptoms.

- **Consequences:** Removal of extraneous details prevents confusion and ensures the report is concise. In the above, the final version of Plan remove the information about online reservation which is not mentioned in the transcript.

PP-6 Necessary statement inclusion

- **Intent and Context:** GPT might miss the information that is mentioned in human reports.
- **Motivation:** Based on the human report, prompts are proposed to GPT to address and supplement any missing information points in the GPT-generated report.

- **Structure and Key Ideas:**

```
1 Please add [X] in [Y]
2 [X]: the needed information
3 [Y]: the place of the information
```

- **Example Implementation:**

```
1 Please add the surgery history of the
   patient in subjective part
2 [X]: the surgery history of the
   patient
3 [Y]: subjective part
```

- **Consequences:** In future iterations, the GPT model will remember to include the patient's surgical history in the subjective part of the report. Utilizing a prompt pattern results in more comprehensive reports. This approach is particularly useful for enhancing the model's ability to construct a fixed structure, making it easier to enumerate the necessary elements.

PP-7 Statement correction

- **Intent and Context:** Using a prompt starting from the beginning differs from using the current prompt based on previous results.
- **Motivation:** The content requires minor modifications after receiving the initial results.

- **Structure and Key Ideas:**

```
1 Please execute [X] based on the above
   version result
2 [X]: repair prompt
```

- **Example Implementation:**

```
1 Please execute omit functional words
   based the above version result.
2 [X]: omit functional words
```

Only use repair prompt Patient reports recurrent ear discomfort starting last Saturday, no recent medical attention for issue. Dissatisfaction with previous treatments including Sofradex ear drops noted. Self-administration of medication four times in past six months yielded favorable results, but not for current episode. Sore throat present, history suggests chronicity.

Using repair prompt with PP-7 Recurrent ear condition since last Saturday, no recent medical attention. Dissatisfied with past treatments, including Sofradex ear drops. Self-treated with unspecified medication, effective four times in six months, ineffective currently. Concurrent sore throat, symptoms suggest chronic issue.

- **Consequences:** In the above example, it is clear that the previous cue includes the subject. This is because one of the prompts was "omit subject" before executing "omit function word". The first result was produced by entering only "omit function word". However, the later result was executed with the subject already omitted.

Using a prompt from the beginning may miss modifications made to the first prompt, focusing only on the later parts. Therefore, not merging two prompts is an effective way to ensure that GPT follows every instruction point in a single prompt. This prompt pattern is aimed to solve the problem that when there are multiple instruction in one repair prompt. It is better to separate them, and execute them one by one to make sure every improvement is based on the previous version. This is also inspired me to create repair prompt in minimum unit.

5.3.3 Report Checking

- PP-8 Requirement validation
 - **Intent and Context:** Verify that GPT adheres to the given prompt instructions, serving as a measure to ensure the completion and accuracy of the response provided by the model.

- **Motivation:** Check the results meet the requirements of GPT itself. Once something goes wrong, it could be modified immediately.

- **Structure and Key Ideas:**

```

1         If [X] does not [Y], then [Z].
2         [X]: the thing with requirements
3         [Y]: requirement
4         [Z]: excution

```

- **Example Implementation:**

```

1         If the assessment is not a clinical word
           or statement, then change the
           assessment into a word or statement.
2         [X]: the assessment
3         [Y]: is not a clinical word or statement
4         [Z]: change the assessment into a
           clinical word or statement.

```

Initial Assessment The patient seems to have had an ear infection which was treated with ear drops, showing improvement until recently. There are no signs of current inflammation, but the ear still has excess skin flakes from the infection. The symptoms could be due to the healing process or the beginning of a new infection.

Second version Assessment after using RP-10 The patient likely experienced acute otitis externa with subsequent resolution; however, the presence of aural fullness or pruritus may be residual effects of the condition or early manifestations of recurrence. There are no overt signs of current infection, but post-inflammatory desquamation is present.

Third version after using the PP-8 Resolved otitis externa with residual cerumen and post-inflammatory debris; no active infection observed. Possibility of subclinical relapse or psychogenic hyperacusis.

- **Consequences:** It has been noted that the application of the PP-8 can potentially enhance the performance of GPT in fulfilling prompts with greater accuracy. This method often yields more favorable results than the reiteration of a repair prompt alone. For instance, in an initial example, the assessment was presented as an extended sentence devoid of clinical terminology. Upon the introduction of a repair prompt, clinical language began to be incorporated into the assessment. However, the response continued to be narrative in style,

which was misaligned with the intent of the repair prompt. The prompt specifically requested a concise clinical word or statement, not an elaborate explanation. By employing the PP-8, the GPT is able to self-evaluate its response to ensure conformity with the stipulated requirements of the prompt. This process facilitates the generation of a final assessment version that succinctly employs clinical terms or statements.

6 Evaluation

This section aims to describe the evaluation of the GPT-generated report. In the common NPL project, ROUGE/BLEU will be used as the common summarization evaluation method. However, ROUGE is compared word by word. The report generated by GPT would usually be longer than a human report. This would make the ROUGE score so low. Thus, this paper explores other evaluation methods.

6.1 Method

The general idea of the evaluation is to compare human reports with GPT-generated reports. Human reports will be regarded as the golden standard. This process includes three steps.

Firstly, split the statement in the human report. Every statement only concludes one information point. This is usually divided by the comma. Also, some abbreviations might represent two pieces of information, such as "ADS redness" which should split into left ear redness and right ear redness. ADS is the abbreviation of the right and left ear in Dutch. The reason for using statements but not words is that words might include functional or other nonetheless words. In the report, the high similarity means the high consistency of the information points. Thus, the evaluation uses a statement as the minimum evaluation unit.

Secondly, identify the corresponding statement in the GPT-generated report. After segmenting the statements, each is located within the GPT-generated report and then classified. Recent research based on doctors' evaluations of GPT-generated reports has divided the types of comparisons into three main categories: Equivalent, Addition, and Error (van Zandvoort et al., 2023). Within these, 'Equivalent' includes 'Identical' and 'Paraphrase'; 'Addition' is split into 'Relevant' and 'Redundant'; 'Error' is further broken down into 'Hallucination', 'Categorization', 'Repetition', 'Omission', and 'Incorrect' (van Zandvoort et al., 2023). The classifications in this study are based on a manual evaluation process, primarily considering the perspective of the GPT-generated report shown in Table 4. The corresponding relationship about classification and these types will be explained in Section 6.2. In contrast, this paper is based on the human

report, observing whether the GPT can produce either identical or thematically similar statements within the same topic. The main difference between these two classification methods is that this paper does not subdivide information in the GPT-generated report that is not in the human report. Instead, it focuses more on how the GPT performs in relation to each statement in the human report.

The last step is to identify the corresponding statement in GPT-generated report belongs to which category according to the definition. Table 5 shows an example of how GPT statements can be categorized.

Classification	Category	Definition
Positive True(TP)	Identical(I)	Information is an exact match to the transcript, including the accurate portrayal of any negations. All the words that contribute to the information point are exactly the same except the functional words.
	Similar(S)	Information closely aligns with the transcript in meaning or content, with negations preserved correctly, though the wording may vary.
False Negative(FN)	Factual Errors(FE)	Information is incorrect or misleading based on the transcripts, including the mishandling of negations that alters the factual content, such as incorrect negation leading to a false hypothetical or historical statement.
	Unsubstantiated Details(UD) Incorrect Statements(IS)	Details are included without support from the transcript. The report contains inaccuracies or distortions of the information compared to human reports.
True negative(TN)	Relevant Additional(RA)	The information that the human report does not cover but corresponds with the transcript.
False Negative(FN)	Omissions(O)	Omission of information point from the human report, potentially leading to an incomplete or misleading medical narrative.

Table 4: Classification and Category Definitions for GPT-Generated Report Statements

Part	Human Report Statement	Human Report Unit	Min	Statement	Max	Category	Corresponding statement in GPT generated report
S	Since 3 days loop ear and pain again now also on the right	Since 3 days loop ear and pain again		now also on the right		Similar	Complaints worsened about 3-4 days ago
	Did see some improvement with antibiotics but lots of side effects from.	AD otorrhea and ear canal redness		Did see some improvement with antibiotics but lots of side effects from.		Similar	Patient reports problems in both ears
O	ADS: otorrhea and ear canal redness	AD otorrhea				Similar	Patient reports problems in both ears
A	Otitis externa on both sides	AS ear canal redness AS otorrhea Right ear Otitis externa				Omission Identical Identical	Right ear is red and fluid is coming out. Left ear is wet and yellow.
P	lom pte	Left ear Otitis externa lom pte				Identical Similar	Otitis externa bilaterally, worse on the left side. Patient prefers hospital2 because of previous negative experience with hospital1, but eventually accepts
	referral ENT	referral ENT				Identical	Referral to ENT specialist at hospital1.

Table 5: Comparison of Human and GPT-Generated Report Statements

6.2 Accuracy Results

This study employs precision and recall as fundamental metrics to evaluate the quality of reports generated by GPT. These metrics typically involve True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). Each category is assigned to one of these types based on its definition. TP denotes instances where the GPT correctly generates content that aligns with the human report, such as identical or similar statements. FP refers to instances where GPT produces information that neither corresponds with the gold standard nor the transcripts. Thus, categories like Factual Errors, Unsubstantiated Details, and Incorrect Statements are considered FP. It is important to note that although TN is usually included in the numerator of precision, in the context of this paper, the focus is on the accuracy based on the human report, not on the accuracy of the GPT-generated content relative to the transcripts. Therefore, TN will not be included in the precision calculation. Since the GPT is generating a report requiring the extraction and correct categorization of information from transcripts, including Relevant Additional would be inappropriate. FN indicates that the GPT has failed to generate information mentioned in the human report, which corresponds to Omissions. Different categories correspond to TP, FP, TN, and FN as shown in Table 6. The formulas for calculating precision and recall are illustrated in Equations 5 and 6.

$$Precision = \frac{TruePositives(TP)}{TotalNumberofStatementsfromHumanReport} \quad (5)$$

$$Recall = \frac{TruePositives(TP)}{TruePositives(TP) + FalseNegatives(FN)} \quad (6)$$

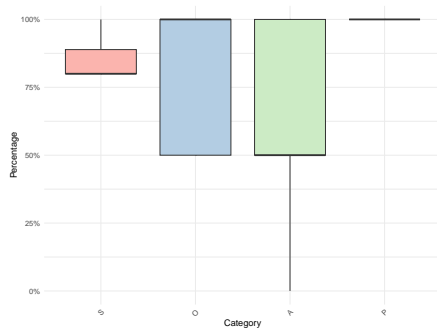


Figure 13: Box Plot of Precision

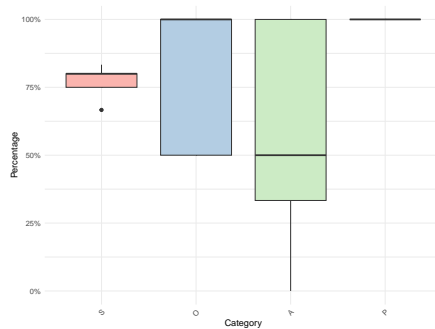


Figure 14: Box Plot of Recall

According to the evaluation methodology outlined, the results of precision and recall assessments based on five datasets is shown in Table 7. And the

box plot of precision and recall are shown in Figure 13 and 14. Precision and recall evaluations were conducted for each section of every report. Overall, the precision rate of the evaluations was 75.92%, with one report achieving as high as 91.67%, while the remaining reports exceeded 70%. This demonstrates that the GPT model mentions over 70% of the information points highlighted by the doctors.

In a detailed analysis of each part, the plan section exhibited the highest accuracy, reaching 100% across all five datasets. This is likely attributable to the fact that doctors' plans are always discussed with patients and thus recorded in the transcripts. Assuming there are no transcription errors, the plan section is expected to be comprehensively captured. The subjective and objective parts followed, with accuracies of 77% and 80%, respectively. These sections' information can also be sourced from recordings; however, the failure to achieve full coverage could be due to the GPT model's incomplete understanding of the content each section should encompass.

Particularly for the subjective section, which consists mostly of patient complaints, the challenge for GPT lies in distilling only the key information relevant to the visit's purpose. GPT tends to include a broader scope of information than necessary and may omit crucial details. Assessment accuracy is primarily determined by the physician's expertise, which explains why this section had the lowest precision in the reports. GPT can only make assessments based on the knowledge it possesses; for instance, Otitis Media with Effusion typically presents without ear pain, in contrast to External Otitis, which does. If the transcript mentions ear pain, GPT might assess the condition as External Otitis, while a doctor could diagnose Otitis Media with Effusion after a comprehensive examination. Thus, GPT's assessment process is limited to information from the transcript, which is substantially less than what a doctor gathers through physical examination, observation, patient interaction, and clinical experience.

Therefore, Assessment is the least accurate section within the reports. However, this trend is not consistent across all reports. In some instances, the precision of the Objective section is lower than that of the Assessment, indicating that GPT's judgment may not be entirely accurate. The model also needs to extract relevant objective parts related to the diagnosis from the transcript. Nonetheless, if the physician's diagnosis does not involve superfluous diagnostic methods, GPT can still provide accurate results. Consequently, some datasets may display varying trends. Information in the Subjective category originates from patient complaints, but the overall tendency is for the precision of the Plan to surpass that of the Objective, which in turn exceeds the Subjective, which is higher than the Assessment.

Analyzing the recall data presented in Table 6, we can observe various insights into the performance of the GPT model across different report sections. The average recall rate for the parts evaluated stands at 81.44%, indicating

a generally high level of model effectiveness in retrieving relevant information points from the transcripts.

In particular, the ‘Plan’ (P) section consistently achieves a recall rate of 100% across all reports, suggesting that the model is highly proficient in identifying and recounting the treatment plans or actions proposed by healthcare professionals. This could be attributed to the structured and explicit manner in which such plans are typically presented within clinical transcripts.

Conversely, the ‘Subjective’ (S) and ‘Objective’ (O) sections exhibit more variability, with recall scores ranging from 50% to 100%. The ‘Objective’ section, which includes observable and measurable facts, achieves perfect recall in three out of the five reports. This might reflect the objective and straightforward nature of the data, which can be directly extracted from the transcripts without significant interpretative variation.

The ‘Assessment’ (A) section, while also recording perfect recall in two reports, shows a significant drop to 0% in one instance, resulting in an average recall of 60%. This variance suggests potential challenges faced by the model in capturing assessments that necessitate a synthesis of disparate information and clinical discernment, which are often not explicitly detailed in the transcripts.

In summary, while GPT demonstrates an adeptness for recalling explicit, clearly defined information from clinical transcripts, it appears there is still a need for improvement in the consistent identification and reporting of more nuanced and complex medical assessments.

Classification	Category
True Positive (TP)	Identical, Similar
False Positive (FP)	Factual Errors, Unsubstantiated Details, Incorrect Statements
True Negative (TN)	Relevant Additionals
False Negative (FN)	Omissions

Table 6: Classification of Categories into TP, FP, TN, and FN

6.3 Medical Summarizing Precision Score(MSPS)

The formula of precision and recall used in the previous section has several limitations. Firstly, the formula of precision and recall does not consider the impact of Factual Errors, Unsubstantiated Details, and Relevant Additions. Due to precision and recall consider the perspective of human reports. This means precision and recall do not consider the information which is not been mentioned in the human reports. Secondly, precision and recall are considered

Precision	Report1	Report2	Report3	Report4	Report5	Avg_part
S	80.00%	66.67%	80.00%	75.00%	83.33%	77.00%
O	50.00%	100.00%	50.00%	100.00%	100.00%	80.00%
A	50.00%	100.00%	50.00%	33.33%	0.00%	46.67%
P	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Avg_Report	70.00%	91.67%	70.00%	77.08%	70.83%	75.92%
Recall	Report1	Report2	Report3	Report4	Report5	Avg_part
S	80.00%	80.00%	100.00%	80.00%	88.89%	85.78%
O	50.00%	100.00%	50.00%	100.00%	100.00%	80.00%
A	100.00%	100.00%	50.00%	50.00%	0.00%	60.00%
P	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
Avg_Report	82.50%	95.00%	75.00%	82.50%	72.22%	81.44%

Table 7: Result of Precision and Recall

equal to identical statements and similar statements. However, even though these two categories are both related to the true positive, there is a quality difference between these two categories.

Thus, the Medical Summarizing Precision Score(MSPS) shown in Equation 6.3 is created to improve the evaluation way on the perspective of the sentence. It consists of three parts.

The first part is the Human Report Reduction Degree(HRRD). In this section, every category that is used to evaluate the precision of the information point mentioned in the human report, namely identical, similar, omission, and incorrect, has a different weight. The weight of each category comes from the experience of comparing GPT reports and human reports when doing the pilot trial. The identical is definitely weight 1. A similar statement usually includes two times more information than a human report, even if they are saying the same thing. GPT is not good at doing simplicity. Similar statements usually contain more details of one information point. This is the reason why the weight of a similar should be less than 1. 0.8 is the general quality compared to an identical statement based on the pilot trial. For instance, GPT may produce extensive passages, as exemplified in the ‘Plan’ section of Table 5. For instance, the term "lom pte" when translated into Dutch as "In overleg met patiënte" means "in consultation with the patient." However, GPT elaborates on the process of consultation between the patient and the doctor at length, which, although classified as ‘similar,’ the ‘identical’ representation would enhance the readability of the report. Omission sometimes refers to those information points that could be concluded by GPT after no more than three prompts. This is the reason why it is not -1 as incorrect. The range of this score is between -1 and 1.

The second part is the Incorrect Additional(IA). This is about the statement that does not have a reference in a human report. The concept of error rate is pivotal in this context, wherein a weight of -1 is ascribed to errors verifiable within the transcript, indicating that GPT has generated incorrect information. For unsubstantiated details lacking evidently support from the transcript, a weight of 0.5 is assigned. The generation of unmentioned information by GPT denotes a negative effect during the report creation process. For a medical report, the inclusion of unverified information is undesirable as it constitutes a negative action. Therefore, a logarithmic function is employed to model this aspect; a higher error rate results in a diminishing value of the function, with the negative impact exponentially increasing. Should the error rate be sufficiently high, it could even reduce the Medical Summarizing Precision Score (MSPS) to zero.

The last section is the score of Relevant Additional(RA). The relevant additional are the information statements that could be found related information in the transcript. The human report does not include this information and cannot totally deny the value of them. This is the reason why this section is positive. However, in the pilot trial, the tendency of GPT to generate sentences that are often verbose is not good for readability. What is worse, the information not mentioned in human reports represents the lower informational priority these information are. Thus, its positive contribution is extremely limited. This is the reason why using the lg function to restrict the value of this part between 0 and 0.1. In general, the maximum of HRRD+IA+RA is 1.1. This is the reason why there is a division in MAPS to make it standardized. Furthermore, the value of AE could be negative infinity. Usually in the evaluation score aspect, the value should not be less than 0. Thus, the MSPS should take the maximum of the sum of these three parts and 0.

$$\begin{aligned}
 MSPS &= \max\left(\frac{HRRD+IA+RA}{1.1}, 0\right), 0 \leq MSPS \leq 1 \\
 \text{where} \\
 HRRD &= \frac{NS - I \times 1 + NS - S \times 0.8 + NS - O \times 0.5 + NS - IS \times (-1)}{NS - HR}, -1 \leq HRRD \leq 1 \\
 IA &= \lg\left[1 - \frac{NS_{FE} \times 1 + NS_{UD} \times 0.5}{NS_{GPTR}}\right], IA \leq 0 \\
 RA &= \lg\left(\frac{NS_{RA} \times 100}{NS_{GPTR}}\right) * 0.1, 0 \leq RA \leq 0.1 \quad (7)
 \end{aligned}$$

6.4 Result of MSPS

The result of using the Medical Summarizing Precision Score is shown in Table 8. Generally, the value of MSPS is lower than that of precision. This is attributed to the formula assigning more weight to inaccurate statements, whether they appear in the human report or the transcripts. Additionally, assigning a lower weight to similar statements also contributes to the reduced value. The trend

where the ‘Plan’ section exhibits the highest value, analogous to precision, persists. A notable distinction, however, is that the values for the ‘Objective’ and ‘Assessment’ sections are almost equally low. This suggests that GPT tends to generate more extraneous information when creating the ‘Objective’ part compared to other sections.

Furthermore, the box plot of MSPS is illustrated in Figure 15. In comparison to the results from the precision and recall metrics, MSPS exhibits a more normal distribution. In the box plots for precision and recall, the distribution of the median in some sections is even identical to the quartile points, indicating a narrower inadequate range.

The MSPS formula offers two significant improvements over precision. First, it provides a sentence-level evaluation, for instance, by differentiating the weights between similar and identical statements. It assesses based on information points formed from the human report, not merely on whether the content is covered, but also considering the presentation of information. Secondly, it is not confined to information addressed in the human report. The formula includes an assessment of information not mentioned in the human report. In summary, MSPS offers a more comprehensive evaluation of GPT-generated reports, representing a more thorough method of assessment.

	Report1	Report2	Report3	Report4	Report5	Avg_part
S	64.14%	60.44%	86.36%	47.18%	87.45%	69.11%
O	72.73%	47.73%	29.08%	76.36%	22.73%	49.73%
A	0.00%	10.52%	38.17%	90.91%	90.92%	46.10%
P	72.73%	90.91%	81.82%	86.36%	92.32%	84.83%
Avg_Report	52.40%	52.40%	58.86%	75.20%	73.35%	62.44%

Table 8: Result of Medical Summarizing Precision Score(MSPS)

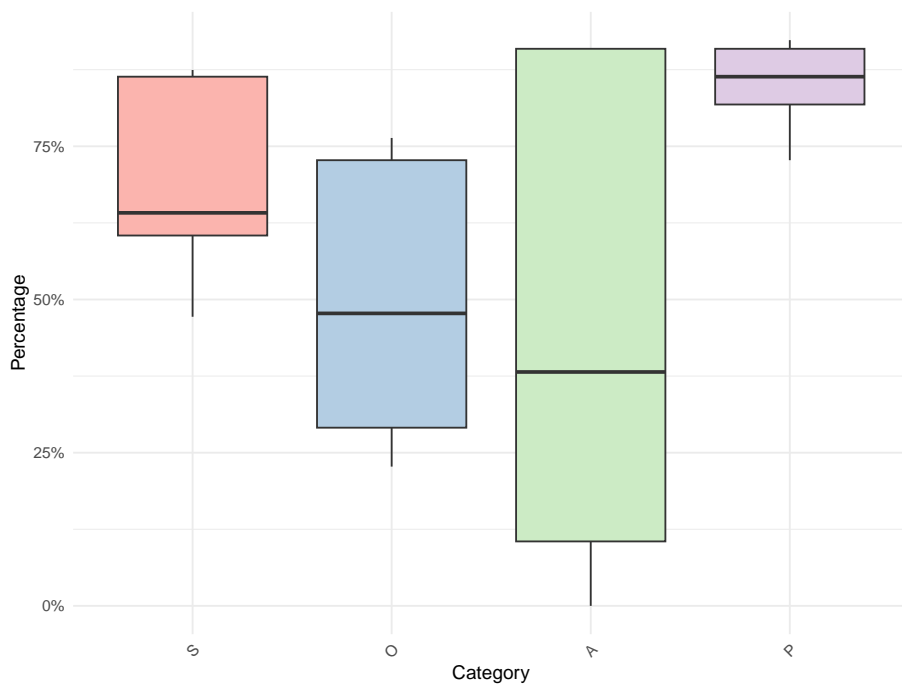


Figure 15: Box Plot of Medical Summarizing Precision Score(MSPS).

7 Treatment Validation

7.1 Conclusion Validity and Internal Validity

The research acknowledges that the limited sample size may pose a constraint. However, this limitation is somewhat offset by the uniform complexity of the sample in Section 4, potentially strengthening the study’s internal validity.

7.2 Construct Validity

To address construct validity, the research introduces the Medical Summarizing Precision Score (MSPS) as an evaluation method, detailed in Section 6. The incorporation of MSPS aims to reduce the effects of measurement validity issues by providing a more nuanced assessment tool.

7.3 External Validity

In combating concerns of generalizability and time sensitivity, the research transcends the boundaries of medical memorization. It innovates by crafting prompt patterns that are not only pertinent to medical texts but also applicable to broader text analysis and categorization domains. These adaptable prompt patterns, encompassing text-related, medical-related, and category-related variations, are intended to be universally applicable, thereby offering a strategic solution to enhance the research’s applicability and longevity.

8 Conclusion & Discussion

8.1 Findings

This study employed prompt engineering to automatically generate SOAP reports conforming to Dutch standards for otolaryngology patient-doctor clinic transcripts. While the initial prompts led to reports significantly different from human-generated ones, iterative improvements aligned the final reports more closely with human reports in terms of information points. However, this doesn't imply that GPT-generated reports are identical to human reports. Due to its inherent logical limitations and the limited diversity of its corpus, GPT struggles with certain nuances, persistently failing to generate accurate information despite multiple prompt attempts.

During the process of generating SOAP reports from transcripts, the study accumulated a substantial collection of prompt statements. These were extracted to derive prompt patterns applicable in medical contexts. These patterns may also be utilized in medical scenarios beyond otolaryngology, marking a significant contribution of this paper.

Additionally, the study evaluated five sample datasets from the research of (van Zandvoort et al., 2023). The precision rate was consistently over 70%, indicating acceptable accuracy of the reports. The average recall rate was 81.44%, suggesting reliability in the content generated by GPT. The assessment focused more on covering the information points in human reports, overlooking the quality of generated information, such as the distinction between identical and similar content. Moreover, precision and recall did not account for incorrect information generated by GPT, even if not mentioned in human reports or transcripts. Therefore, this paper proposes the Medical Summarizing Precision Score (MSPS), a new evaluation method that differentiates more finely between identical and similar sentences and amplifies the impact of errors on the overall report value. Under MSPS, the same dataset scored 62.44%. Although lower than the precision rate, it still exceeds 60% and shows a converging trend with the distribution of scores generated by precision. This indicates that the new evaluation method aligns with data patterns and can be applied in subsequent similar research.

8.2 Limitation

Limited Sample Size: This study, while contributing valuable insights into the application of GPT for medical report summarization, has several limitations. First, the sample size used for evaluation is not extensive. The findings presented are based on a limited dataset, which may not fully capture the variability inherent in broader medical practice.

Generalizability: Secondly, the study’s focus on otitis media consultations constrains its generalizability. The applicability of the Medical Summarizing Precision Score (MSPS) in other medical scenarios remains to be confirmed, and further research is required to ascertain its effectiveness across diverse medical contexts.

Reliance on Transcripts: Furthermore, GPT’s reliance solely on transcripts for generating reports is a notable limitation. This approach lacks the nuanced clinical insights that come from physical diagnosis and other physician-specific diagnostic methods, which are integral to the practice of medicine. The absence of these elements in the AI-generated reports calls into question the depth and clinical accuracy of the findings.

Human Evaluation Time-consuming: Lastly, the process of human evaluation, while necessary for the validation of the model’s performance, is time-consuming and labor-intensive. The reliance on human evaluators presents challenges in terms of scalability and efficiency, which are critical for the practical application of such AI systems in healthcare.

8.3 Future work

The findings of this study lay the groundwork for several avenues of future research. To address the limitation of sample size, subsequent studies should employ a larger sample that encompasses a wider array of medical conditions beyond otitis media. This expansion would not only enhance the robustness of the results but also validate the model’s applicability across diverse medical scenarios.

In the development of the GPT model, future iterations could benefit from training on a comprehensive medical text corpus within the GPT Explore framework. Such a corpus would include a broader spectrum of medical literature, ensuring that the model is well-informed by a vast array of clinical knowledge and terminologies.

Another important direction for future work is the systematization of category recognition within the GPT model. By refining the model’s ability to categorize information accurately, we can significantly reduce the current reliance on time-intensive human evaluation. This would involve developing more sophisticated algorithms capable of discerning nuances in medical data, thereby streamlining the validation process.

Collectively, these efforts will pave the way for creating more autonomous, reliable, and efficient AI systems in healthcare, ultimately advancing the field of medical AI towards more practical and widespread clinical application.

References

- Agrawal, M., Hegselmann, S., Lang, H., Kim, Y., & Sontag, D. (2022). Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 conference on empirical methods in natural language processing* (pp. 1998–2022).
- Alsentzer, E., & Kim, A. (2018). Extractive summarization of ehr discharge notes. *arXiv preprint arXiv:1810.12085*.
- Arndt, B. G., Beasley, J. W., Watkinson, M. D., Temte, J. L., Tuan, W.-J., Sinsky, C. A., & Gilchrist, V. J. (2017). Tethered to the ehr: primary care physician workload assessment using ehr event log data and time-motion observations. *The Annals of Family Medicine*, 15(5), 419–426.
- Brenneman, L. (2001). Guidelines for writing soap notes and history and physicals. *Glen Gardner, NJ: NPCEU Inc*, 26.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., . . . others (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877–1901.
- Cameron, S., & Turtle-Song, I. (2002). Learning to write case notes using the soap format. *Journal of Counseling & Development*, 80(3), 286–292.
- Campanella, P., Lovato, E., Marone, C., Fallacara, L., Mancuso, A., Ricciardi, W., & Specchia, M. L. (2016). The impact of electronic health records on healthcare quality: a systematic review and meta-analysis. *The European Journal of Public Health*, 26(1), 60–64.
- Chintagunta, B., Katariya, N., Amatriain, X., & Kannan, A. (2021). Medically aware gpt-3 as a data generator for medical dialogue summarization. In *Machine learning for healthcare conference* (pp. 354–372).
- Chiu, C.-C., Tripathi, A., Chou, K., Co, C., Jaitly, N., Jaunzeikare, D., . . . others (2017). Speech recognition for medical conversations. *arXiv preprint arXiv:1711.07274*.
- Dai, H., Liu, Z., Liao, W., Huang, X., Wu, Z., Zhao, L., . . . others (2023). Chataug: Leveraging chatgpt for text data augmentation. *arXiv preprint arXiv:2302.13007*.
- Earls, S. T., Savageau, J. A., Begley, S., Saver, B. G., Sullivan, K., & Chuman, A. (2017). Can scribes boost fps’ efficiency and job satisfaction.
- Elmore, R. F. (1991). Comment on “towards rigor in reviews of multivocal literatures: Applying the exploratory case study method”. *Review of educational research*, 61(3), 293–297.
- Falcetta, F. S., de Almeida, F. K., Lemos, J. C. S., Goldim, J. R., & da Costa, C. A. (2023). Automatic documentation of professional health interactions: A systematic review. *Artificial Intelligence in Medicine*, 102487.
- Feng, Z., Guo, D., Tang, D., Duan, N., Feng, X., Gong, M., . . . others (2020). Codebert: A pre-trained model for programming and natural languages. *arXiv preprint arXiv:2002.08155*.
- Finley, G., Edwards, E., Robinson, A., Brenndoerfer, M., Sadoughi, N., Fone, J., . . . Suendermann-Oeft, D. (2018). An automated medical scribe for documenting clinical encounters. In *Proceedings of the 2018 conference of*

- the north american chapter of the association for computational linguistics: Demonstrations* (pp. 11–15).
- Garousi, V., Felderer, M., & Mäntylä, M. V. (2019). Guidelines for including grey literature and conducting multivocal literature reviews in software engineering. *Information and software technology, 106*, 101–121.
- Giorgi, J., Toma, A., Xie, R., Chen, S., An, K. R., Zheng, G. X., & Wang, B. (2023). Clinical note generation from doctor-patient conversations using large language models: Insights from mediq-chat. *arXiv preprint arXiv:2305.02220*.
- Golob Jr, J. F., Como, J. J., & Claridge, J. A. (2016). The painful truth: The documentation burden of a trauma surgeon. *Journal of Trauma and Acute Care Surgery, 80*(5), 742–747.
- Gutiérrez, B. J., McNeal, N., Washington, C., Chen, Y., Li, L., Sun, H., & Su, Y. (2022). Thinking about gpt-3 in-context learning for biomedical ie? think again. *arXiv preprint arXiv:2203.08410*.
- Heart, T., Ben-Assuli, O., & Shabtai, I. (2017). A review of phr, emr and ehr integration: A more personalized healthcare and public health policy. *Health Policy and Technology, 6*(1), 20-25. Retrieved from <https://www.sciencedirect.com/science/article/pii/S2211883716300624> doi: <https://doi.org/10.1016/j.hlpt.2016.08.002>
- Hu, M., Pan, S., Li, Y., & Yang, X. (2023). Advancing medical imaging with language models: A journey from n-grams to chatgpt. *arXiv preprint arXiv:2304.04920*.
- Jacsó, P. (2005). Google scholar: the pros and the cons. *Online information review*.
- Jha, A. K., Doolan, D., Grandt, D., Scott, T., & Bates, D. W. (2008). The use of health information technology in seven nations. *International journal of medical informatics, 77*(12), 848–854.
- Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., ... Wang, Y. (2017). Artificial intelligence in healthcare: past, present and future. *Stroke and vascular neurology, 2*(4).
- Knoll, T., Moramarco, F., Papadopoulos Korfiatis, A., Young, R., Ruffini, C., Perera, M., ... Savkov, A. (2022, July). User-driven research of medical note generation software. In *Proceedings of the 2022 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 385–394). Seattle, United States: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2022.naacl-main.29> doi: 10.18653/v1/2022.naacl-main.29
- Li, X. L., & Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.
- Lin, C.-Y. (2004, July). ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out* (pp. 74–81). Barcelona, Spain: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/W04-1013>

- Liu, C., Wang, P., Xu, J., Li, Z., & Ye, J. (2019). Automatic dialogue summary generation for customer service. In *Proceedings of the 25th acm sigkdd international conference on knowledge discovery & data mining* (pp. 1957–1965).
- Liu, P., Qiu, X., & Huang, X. (2016). Recurrent neural network for text classification with multi-task learning. *arXiv preprint arXiv:1605.05101*.
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9), 1–35.
- Liu, P. J., Saleh, M., Pot, E., Goodrich, B., Sepassi, R., Kaiser, L., & Shazeer, N. (2018). Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*.
- Liu, Z., Yu, X., Zhang, L., Wu, Z., Cao, C., Dai, H., ... others (2023). Deid-gpt: Zero-shot medical text de-identification by gpt-4. *arXiv preprint arXiv:2303.11032*.
- Maas, L., Geurtsen, M., Nouwt, F., Schouten, S., Van De Water, R., Van Dulmen, S., ... Brinkkemper, S. (2020). The care2report system: Automated medical reporting as an integrated solution to reduce administrative burden in healthcare. In *Hicss* (pp. 1–10).
- Maas, L., Kisjes, A., Hashemi, I., Heijmans, F., Dalpiaz, F., Van Dulmen, S., & Brinkkemper, S. (2021). Automated medical reporting: From multimodal inputs to medical reports through knowledge graphs. In *Healthinf* (pp. 509–514).
- March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision support systems*, 15(4), 251–266.
- Moramarco, F., Papadopoulos Korfiatis, A., Perera, M., Juric, D., Flann, J., Reiter, E., ... Savkov, A. (2022, May). Human evaluation and correlation with automatic metrics in consultation note generation. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 5739–5754). Dublin, Ireland: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2022.acl-long.394> doi: 10.18653/v1/2022.acl-long.394
- Nori, H., King, N., McKinney, S. M., Carignan, D., & Horvitz, E. (2023). Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.
- Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 311–318).
- Rao, S. K., Kimball, A. B., Lehrhoff, S. R., Hidrue, M. K., Colton, D. G., Ferris, T. G., & Torchiana, D. F. (2017). The impact of administrative burden on academic physicians: results of a hospital-wide physician survey. *Academic Medicine*, 92(2), 237–243.
- Reynolds, L., & McDonell, K. (2021). Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended abstracts of the 2021*

- chi conference on human factors in computing systems* (pp. 1–7).
- Sarker, S., Qian, L., & Dong, X. (2023). Medical data augmentation via chatgpt: A case study on medication identification and medication event classification. *arXiv preprint arXiv:2306.07297*.
- Schick, T., & Schütze, H. (2020). Exploiting cloze questions for few shot text classification and natural language inference. *arXiv preprint arXiv:2001.07676*.
- Schukow, C., Smith, S. C., Landgrebe, E., Parasuraman, S., Folaranmi, O. O., Paner, G. P., & Amin, M. B. (2023). Application of chatgpt in routine diagnostic pathology: Promises, pitfalls, and potential future directions. *Advances in anatomic pathology*, 10–1097.
- Sivarajkumar, S., & Wang, Y. (2022). Healthprompt: A zero-shot learning paradigm for clinical natural language processing. In *Amia annual symposium proceedings* (Vol. 2022, p. 972).
- Smith, K., Megyesi, B., Velupillai, S., & Kvist, M. (2014). Professional language in swedish clinical text: Linguistic characterization and comparative studies. *Nordic Journal of Linguistics*, 37(2), 297–323.
- Sorensen, T., Robinson, J., Rytting, C. M., Shaw, A. G., Rogers, K. J., DeLorey, A. P., ... Wingate, D. (2022). An information-theoretic approach to prompt engineering without ground truth labels. *arXiv preprint arXiv:2203.11364*.
- Srivastava, A., Rastogi, A., Rao, A., Shoeb, A. A. M., Abid, A., Fisch, A., ... others (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.
- Tom, E., Aurum, A., & Vidgen, R. (2013). An exploration of technical debt. *Journal of Systems and Software*, 86(6), 1498–1516.
- van Zandvoort, D., Wiersema, L., Huibers, T., van Dulmen, S., & Brinkkemper, S. (2023). Enhancing summarization performance through transformer-based prompt engineering in automated medical reporting. *arXiv preprint arXiv:2311.13274*.
- Wang, J., Shi, E., Yu, S., Wu, Z., Ma, C., Dai, H., ... others (2023). Prompt engineering for healthcare: Methodologies and applications. *arXiv preprint arXiv:2304.14670*.
- Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., ... Le, Q. V. (2021). Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.
- White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., ... Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.
- Wieringa, R. J. (2014). *Design science methodology for information systems and software engineering*. Springer.
- Woolhandler, S., & Himmelstein, D. U. (2014). Administrative work consumes one-sixth of us physicians' working hours and lowers their career satisfaction. *International Journal of Health Services*, 44(4), 635–642.
- Wu, S., Roberts, K., Datta, S., Du, J., Ji, Z., Si, Y., ... others (2020). Deep learning in clinical natural language processing: a methodical re-

view. *Journal of the American Medical Informatics Association*, 27(3), 457–470.

Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., & Ba, J. (2022). Large language models are human-level prompt engineers. *arXiv preprint arXiv:2211.01910*.

A PDD Activity Table

Table 9 is the activity table for PDD.

Activity	Sub-Activity	Description
Data Collection	Request for the data	Initiating the data collection process by submitting a request form.
	Collect the data	Gathering necessary data based on the initial request.
	Create BASIC PROMPT	Developing BASIC PROMPT from the collected data by the researcher.
Generate SOAP Report	Arrange for medical reporting pilot	Setting up the environment for the generation of the SOAP report.
	Input BASIC PROMPT	Input the BASIC PROMPT to GPT as one of the content to generate the SOAP medical report.
Improve SOAP Report	Input TRANSCRIPT	Input the TRANSCRIPT to GPT to generate the SOAP medical report.
	Evaluate DRAFT SOAP	Comparing the DRAFT SOAP against HUMAN REPORT to assess the accuracy.
	Create REPAIR PROMPT	Generating a REPAIR PROMPT to address any needed corrections identified in the evaluation.
	Input REPAIR PROMPT	Implementing the REPAIR PROMPT to finalize and refine the SOAP report.

Table 9: Activity Table

B PDD Concept Table

Table 10 is the concept table for PDD.

Concept	Description
TICKET	Represents the initial request for data, including a unique number(Id), request content and time.
TRANSCRIPTS SET	Detailed data set to be included in the report, containing Id, TRANSCRIPT and HUMAN REPORT.
BASIC PROMPT DRAFT SOAP	Initial prompt to initiate the SOAP report creation. Preliminary version of the SOAP report, including SUBJECTIVE DRAFT, OBJECTIVE DRAFT, ASSESSMENT DRAFT and PLAN DRAFT
HUMAN EVALUATION	A reference report used for comparison to evaluate the quality and accuracy of the SOAP draft, compared with HUMAN REPORT.
REPAIR PROMPT	A prompt created to correct and improve the SOAP DRAFT based on the evaluation, which also includes type and version. Type is used to categorize repair prompt according their functions. Id is used to give the unique number of each REPAIR PROMPT.
FINAL SOAP	The completed and corrected SOAP report, which contains the finalised components after input of repair prompts.

Table 10: Concept Table

C Example1 used in PP-1

S: Decreased hearing and sensitivity li/re. Nose drops and blowing out no effect.

O: AD : redness and swelling of ear canal and redness and scaling of auricle. AS
: redness of ear canal.

A: Otitis externa bilateral

P: Sofradex 2 dd 2 drops, rub skin with cream, revision 1 week

D Guideline1 in used in PP-2

Firstly, be brief and concise, keep quotes to a minimum, use an active voice, use precise and descriptive terms, start each new entry with date and time of session, document all contacts or attempted contacts. Secondly, avoid using names of other clients, family members, or others named by client. Avoid terms like seems, appears. Avoid using the term "relevant", the preferred term is "secondary". Avoid the term "suffer". Avoid value-laden language, common labels, opinionated statements. Do not use terminology unless trained to do so. Thirdly, if the exact nature of the problem is unclear, use the R/O diagnosis or "EU" (for "etiology unknown")i.e., "herpetic lesion EU; R/O contact phlebitis. The term "cannot be excluded" can also be used, i.e. "migraine cannot exclude an aneurysm" or "cough cannot exclude pneumonia".