

Enhancing Summarization Performance Through Transformer-Based Prompt Engineering in Automated Medical Reporting

Daphne van Zandvoort¹ ^a, Laura Wiersema¹ ^b, Tom Huibers², Sandra van Dulmen³ ^c,
and Sjaak Brinkkemper¹ ^d

¹Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands

²Verticai, Utrecht, The Netherlands

³Nivel (Netherlands Institute for Health Services Research), Utrecht, Netherlands

Keywords: Prompt Engineering, Automated Medical Reporting, Medical Dialogue Summarization, SOAP Reporting, Performance, Care2Report.


Abstract: Customized medical prompts enable Large Language Models (LLM) to effectively address medical dialogue summarization. The process of medical reporting is often time-consuming for healthcare professionals. Implementing medical dialogue summarization techniques presents a viable solution to alleviate this time constraint by generating automated medical reports. The effectiveness of LLMs in this process is significantly influenced by the formulation of the prompt, which plays a crucial role in determining the quality and relevance of the generated reports. In this research, we used a combination of two distinct prompting strategies, known as *shot prompting* and *pattern prompting* to enhance the performance of automated medical reporting. The evaluation of the automated medical reports is carried out using the ROUGE score and a human evaluation with the help of an expert panel. The two-shot prompting approach in combination with scope and domain context outperforms other methods and achieves the highest score when compared to the human reference set by a general practitioner. However, the automated reports are approximately twice as long as the human references, due to the addition of both redundant and relevant statements that are added to the report.


1 INTRODUCTION


The application of Artificial Intelligence (AI), notably Machine Learning (ML), to enhance healthcare and assist medical decision-making is a rapidly growing field (Hicks et al., 2022). Large Language Models (LLM) are effectively tackling challenging healthcare tasks, such as disease diagnosis, treatment planning, and medical reporting, using personalized medical prompts, even with limited data (Wang et al., 2023). Prompt engineering in the medical domain, including classification, data generation, anomaly detection, content augmentation, question answering, and medical inference, is crucial in improving these healthcare outcomes (Wang et al., 2023). Ensuring high levels of accuracy and reliability in these AI-driven healthcare applications is essential for their successful inte-


gration into medical support systems (Balagurunathan et al., 2021).

Expanding on the role of AI and ML in healthcare, Electronic Health Records (EHRs) have become a pivotal focus, revolutionizing medical data management and communication (Coorevits et al., 2013). EHR documentation has led to significant changes in medical practice with an increase in data access and communication among medical professionals compared to paper records (Overhage and McCallie Jr, 2020). However, one significant challenge has been the time-consuming data input and hindrances to in-person patient care, resulting in professional dissatisfaction (Friedberg et al., 2014). In response, to lessen this administrative burden, automation of this process was developed by several research initiatives, as demonstrated with the Systematic Literature Review of (van Buchem et al., 2021). Care2Report (C2R) is the only scientific initiative that focuses on the Dutch medical field and automates medical reporting by utilizing multimodal consultation recordings (audio, video, and Bluetooth), enabling knowl-

^a  <https://orcid.org/0009-0003-7143-6696>

^b  <https://orcid.org/0009-0007-8621-3611>

^c  <https://orcid.org/0000-0002-1651-7544>

^d  <https://orcid.org/0000-0002-2977-8911>

edge representation, ontological dialogue interpretation, report production, and seamless integration with electronic medical record systems (Maas et al., 2020; ElAssy et al., 2022). This automated medical reporting serves as a prime example of prompt engineering, specifically in the domain of medical dialogue summarization (MDS), illustrating how technology can streamline healthcare processes.

In automated MDS, the generation of automated medical reporting relies on utilizing state-of-the-art LLMs like Generative Pre-trained Transformers (GPT). The level of detail and specificity in the prompts directly influences the model's comprehension and its ability to produce the expected results (Bigelow, 2023; Heston, 2023; Robinson, 2023). Several articles on the effective crafting of prompts, emphasize the significance of context and clarity in the prompts, including the provision of additional relevant information for optimal results (Bigelow, 2023; Robinson, 2023).

Although prompt engineering has a substantial impact on the performance of LLMs, its full potential in the domain of medical problem-solving remains largely unexplored. Thus, this research aims to answer the following research question:

RQ: Which Prompt Formulation Detail Yields High Performance in Automated Medical Reporting?

To answer the research question we focus on prompt engineering related to automated medical reporting. First, we reviewed existing literature for research within prompt engineering, automatic text summarization, and medical dialogue summarization (Section 2). Subsequently, Section 3 reports on prompt formulation, execution, and analysis. The findings are presented and discussed (Section 4). Finally, the work is summarized and suggestions are provided for future work (Section 5).

2 RELATED WORK

This study builds on prior research in the realm of prompt engineering, aiming to employ diverse prompting methodologies for generating automated medical reports within MDS, a subset of Automatic Text Summarization (ATS).

2.1 Prompt Engineering

A human-initiated prompt serves as the initial step for GPT in comprehending the context and meeting user expectations by producing the desired output (White

et al., 2023). This process includes designing, implementing, and refining prompts to optimize their efficacy in eliciting this intended result (Heston, 2023). An example prompt in the context of this work is shown in Listing 1.

```
1 You are a bot that generates a medical
  report in SOAP format based on a
  conversation between a doctor and a
  patient.
2 Only extract information from the
  conversation to produce the
  Subjective, Objective, Analysis, and
  Plan sections of the medical report
```

Listing 1: Example of a prompt for automated medical reporting. The example is based on existing research of the C2R program.

Based on literature, we decided to use the shot prompting and pattern prompting methods to achieve the highest-performing output since these provide an opportunity to demonstrate an example of the expected output and to delineate the context.

2.1.1 Shot Prompting

In-context learning is a method where language models learn tasks through a few examples provided as demonstrations (Dong et al., 2022). *Shot prompting* employs in-context learning to guide the model's output. There are three strategies: zero-shot, one-shot, and few-shot prompting (Anil, 2023). *Zero-shot prompting*, also known as direct prompting, involves giving the model a task without specific examples, relying solely on the knowledge acquired during training (Anil, 2023). In contrast, *one-shot* and *few-shot prompting* provide examples or 'shots' to the model at run-time, serving as references for the expected response's structure or context (Anil, 2023; Reynolds and McDonell, 2021). The model then infers from these examples to perform the task. Since examples are presented in natural language, they provide an accessible way to engage with language models and facilitate the incorporation of human knowledge into these models through demonstrations and templates (Brown et al., 2020; Dong et al., 2022; Liu et al., 2023a). Currently, there is no universally standardized methodology for providing examples in shot-prompting (Anil, 2023; Dragon, 2023; Tam, 2023). For more straightforward tasks, like language translation or classification, a prompt could be formulated as demonstrated in Listing 2. For more complex tasks, like content generation, a prompt can be constructed as demonstrated in Listing 3.

```
1 Text: My ear feels fine after the
  treatment.
2 Classification: Positive
```

```

3 Text: The doctor examined my ear, and
  everything seems normal.
4 Classification: Neutral
5 Text: I experience some discomfort, I
  suspect it might be ear infection.
6 Classification: Negative
7 Text: The ear pain is unbearable, I
  need to see a specialist.
8 Classification:

```

Listing 2: Example of few-shot prompt in a straightforward task.

```

1 Write a medical report about the
  following transcript: [transcript].
2 Use the following SOAP reports [
  example report 1] and [example
  report 2] as a guide.

```

Listing 3: Example of few-shot prompt in a complex task.

2.1.2 Pattern Prompting

Pattern prompting involves the availability of various patterns that can be chosen and employed as the basis for the formulation of prompts. These patterns facilitate interactions with conversational LLMs across various contexts, extending beyond just discussing interesting examples or domain-specific prompts (White et al., 2023). The aim is to codify this knowledge into pattern structures that enhance the ability to apply it in different contexts and domains where users encounter similar challenges, although not necessarily identical ones. This approach promotes greater reuse and adaptability of these patterns for diverse use cases and situations (White et al., 2023).

The study of (White et al., 2023) introduces, among others, the *context control pattern* category. Context control captures the *context manager pattern*, which enables users to specify or remove context from the prompt. “By focusing on explicit contextual statements or removing irrelevant statements, users can help the LLM better understand the question and generate more accurate responses” (White et al., 2023). The greater the clarity in the statements, the higher the likelihood that the LLM will respond with the intended action. Possible context statements are: “within the scope of X”, “consider Y”, “ignore Z”; an example is shown in Listing 4 (White et al., 2023).

```

1 Listen to this transcript between
  doctor and patient and make a EHR
  entry from it.
2 Consider the medical guidelines.
3 Do not consider irrelevant statements.

```

Listing 4: Example of a prompt using the context manager pattern.

2.2 Automatic Text Summarization

Since the introduction of transformer-based methods in ATS, the usage of prompt engineering has been instrumental in enhancing the performance of ATS processes. In ATS, various pragmatic algorithms can be integrated into computers to generate concise summaries of information (Mridha et al., 2021). When used in Natural Language Processing (NLP), ATS is used to evaluate, comprehend, and extract information from human language (Mridha et al., 2021). The introduction of transformer-based models like GPT (Radford et al., 2019) shows improved performance in NLP-tasks (Mridha et al., 2021) which is beneficial for abstractive summarization.

Abstractive summarization creates summaries by introducing new phrases or words not present in the original text. To achieve accurate abstractive summaries, the model must thoroughly comprehend the document and express that comprehension concisely through new terms or alternative expressions (Widyassari et al., 2019). The opposite of abstractive summarization, is extractive summarization, a method where the summary consists entirely of extracted content (Widyassari et al., 2019). Extractive summarization has been used most frequently because it is easier, but the summaries generated are far from human-made summaries, in contrast to abstractive summarization (Widyassari et al., 2019; Yadav et al., 2022).

2.3 Medical Dialogue Summarization

In MDS, it is important that the summaries are at least partly abstractive. In one respect, the reports are generated from dialogue, so extracting literal (sub-)sentences will not lead to a coherent report; conversely, the summaries must be comparable to the human-made versions of the general practitioners (GP). In MDS, the relevant medical facts, information, symptoms, and diagnosis must be retrieved from the dialogue and presented either in the form of structured notes or unstructured summaries (Jain et al., 2022). The most common type of medical notes are SOAP notes: **S**ubjective information reported by the patient, **O**bjective observations, **A**ssessment by medical professional and future **P**lans (Jain et al., 2022; Krishna et al., 2021).

Previous work in MDS has produced the transformer-based approaches of MEDSUM-ENT (Nair et al., 2023), MedicalSum (Michalopoulos et al., 2022), and SummQA (Mathur et al., 2023).

Table 1: Example of part of a consultation transcript and the corresponding SOAP report (translated to English, the original transcript and SOAP report are in Dutch).

Transcript	SOAP report
<p>GP: Good morning. P: Good morning, hello. Last week I visited your colleague. GP: Yes I see, for your ear. P: I had an ear infection. Well, I'm actually getting sicker. Since yesterday, I've been getting sicker and sicker. GP: She gave you antibiotics, right? P: Yes, the first three or four tablets were really like, whoa. And after that, it was just the same. So, I still have ear pain. And now I notice that my resistance is decreasing because of the antibiotics. I'm just getting more tired now. ... GP: We're just going to take a look. There is some fluid. Also, air bubbles behind the eardrum. That is clearly visible. P: Yes, yes, that's correct. It gurgles and it rattles and it rings. And it's just blocked. GP: Yes, I believe that when I see it like this. It doesn't look red. It doesn't appear to be really inflamed. ... GP: I think, for now, at least, you should finish the antibiotics. P: That's two more days. GP: Yes, and continue using the nasal spray, or the other nasal spray, for another week and see how it goes. Just come back if it's still not better after a week. And if it persists, well, maybe then you should see the ENT specialist. ... GP = General Practitioner, P = Patient</p>	<p>S: Since 1.5 weeks, ear pain and a feeling of deafness right ear, received antibiotics from the GP. Feeling sicker since yesterday, experiencing many side effects from the antibiotics. Using Rhinocort daily for hyperreactivity. Left ear operated for cholesteatoma, no complaints. O: right ear: eardrum visible, air bubbles visible, no signs of infection. [left ear ?] A: OMA right P: Advice xylomethazine 1 wk, continue antibiotics, review symptoms in 1 week. Consider prescribing Flixonase, referral to ENT?</p> <p>OMA = Otitis Media Externa, ENT = Ear, Nose, Throat</p>

- “MEDSUM-ENT is a medical conversation summarization model that takes a multi-stage approach to summarization, using GPT-3 as the backbone”. MEDSUM-ENT first extracts medical entities and their affirmations and then includes these extractions as additional input that informs the final summarization step through prompt chaining. Additionally, MEDSUM-ENT exploits few-shot prompting for medical concept extraction and summarization through in-context example selection. Their study concludes that summaries generated using this approach are clinically accurate and preferable to naive zero-shot summarization with GPT-3 (Nair et al., 2023).
- MedicalSum is a sequence-to-sequence architecture for summarizing medical conversations by integrating medical domain knowledge from the Unified Medical Language System (UMLS) to increase the likelihood of relevant medical facts being included in the summarized output. Their analysis shows that MedicalSum produces accurate AI-generated medical documentation (Michalopoulos et al., 2022).
- SummQA is a “two-stage process of selecting semantically similar dialogues and using the top-k similar dialogues as in-context examples for GPT-4”. They generate section-wise summaries and classify these summaries into appropriate section

headers. Their results highlight the effectiveness of few-shot prompting for this task (Mathur et al., 2023).

The present study not only builds upon this existing knowledge base by integrating a combination of shot prompting and context patterns into prompt engineering but also includes a crucial human evaluation component, in addition to the accuracy measurement. This human evaluation provides comprehensive insights into prompt performance beyond computer-based metrics. Leveraging GPT-4 for Dutch consultations, we ensure that the resulting medical reports adhere to the widely recognized SOAP guidelines. It is noteworthy that while prior studies have demonstrated the efficacy of shot-prompting, this is the first published study to incorporate both shot prompting and context pattern prompting in the domain of Dutch MDS, thereby making a significant contribution to the Dutch medical field.

3 STUDY DESIGN

We conducted a causal-comparative study to identify the cause-effect relationship between the formulation detail of the prompt and the performance of the automated medical report (Schenker and Rumrill Jr, 2004). We followed the approach of the C2R pro-

gram, by using transcripts that were made of the verbal interaction during a series of video-recorded consultations between GPs and their patients (Figure 1) (Maas et al., 2020; Meijers et al., 2019). The recordings, for which patients as well as GPs provided informed consent, were made as part of previous communication projects carried out by researchers at Radboudumc and Nivel (Netherlands institute for health services research) (Houwen et al., 2017). Subsequently, medical professionals examined these transcripts to generate SOAP medical reports, with an illustrative example presented in Table 1. These SOAP reports are used in the study as a human reference for comparison with the automatically generated reports. The automatically generated medical reports were produced by GPT based on various prompt formulations. Using prompt engineering, the prompts were created using the *shot prompting* and *context manager pattern* techniques. Each executed prompt resulted in medical reports that were analyzed to determine which prompt yielded the best results.

3.1 Formulation of Prompts

The prompts formulated in this work combine *shot prompting* and *context pattern prompting* (Figure 2). First, a base prompt was established upon which all other elements in the prompt could be built. Variability in performance can then be attributed solely to differences in shots or context, rather than possible other factors. The base prompt compels the GPT to solely utilize elements present in the transcript to

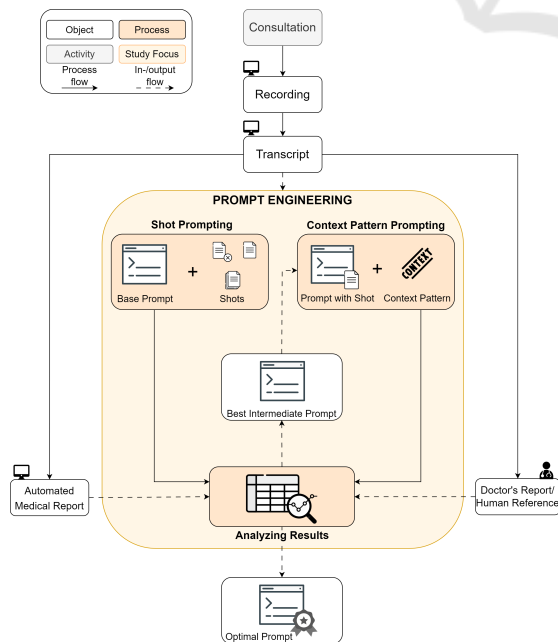


Figure 1: Research method visualization.

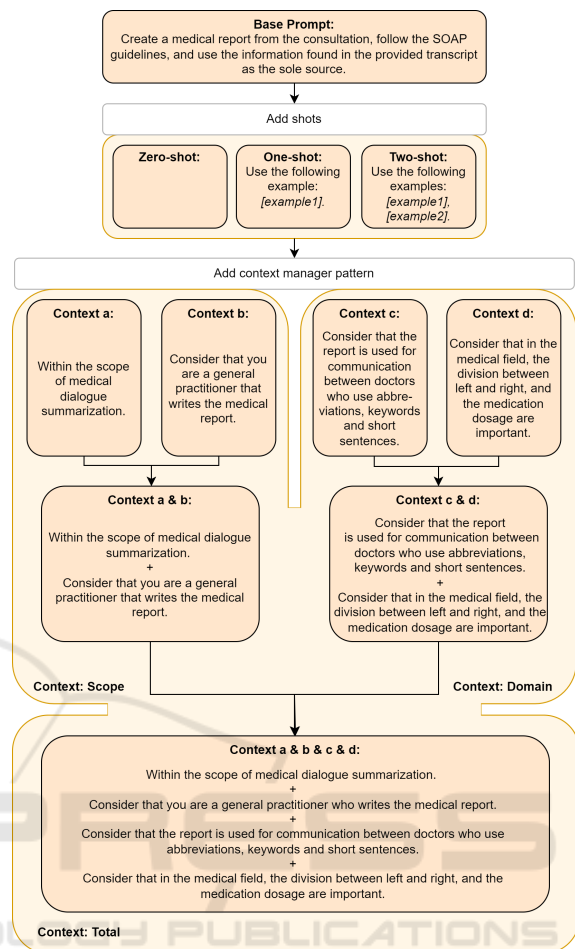


Figure 2: The flow of prompt formulation (translated to English, the original prompts are in Dutch).

prevent hallucinations (Banerjee et al., 2023; Ji et al., 2023).

This base prompt was initially employed to construct three versions of *shot-prompting*: zero-shot, one-shot, and two-shot. The most effective shot-prompting among these three prompts was selected. Using the *context manager pattern*, an increase in context was added to the prompt to measure the effect of incorporating more context into the prompt. The context is divided into two types of contexts: scope context and domain context. The scope context explains in what scope the GPT operates and what its role is. The domain context gives more details about communication and important elements in the medical field.

The following context statements are included:

- Within the scope of medical dialogue summarization;
- Consider that you are a general practitioner who writes the medical report during the consultation;

- c. Consider that the report is used for communication between doctors who use abbreviations and short sentences or keywords;
- d. Consider that in the medical field, the division between left and right, and the medication dosage are important.

Each of these statements, as well as various combinations of them, as illustrated in Figure 2, were included to assess their individual, as well as their combined effects.

3.2 Running of Prompts

The crafted prompts served as a means to collect and assess the data concerning their performance in practice. The formulated prompts were run in a self-written prompt engineering software supported by the Azure OpenAI Service, which is a fully managed service that allows developers to easily integrate OpenAI models into applications (Mickey, 2023). GPT-4 was used with a temperature of 0; GPT-4 is the current best-performing GPT-version and a temperature of 0 minimizes the creativity and diversity of the text generated by the GPT model (Liu et al., 2023b).

As a data source, seven real-world Dutch consultations between a GP and patients, concerning Otitis Externa and Otitis Media Acuta, were utilized and employed in three distinct manners:

- Five transcriptions of these consultations served as input data to create automated medical reports.
- Five manually created SOAP reports (of these five transcriptions) by doctors were employed as a human reference for the automated medical reports.
- Two manually created SOAP reports by doctors were used as examples in shot-prompting.

It has been ensured that both an external and middle ear infection consultation are included in the examples, but the distinction between input and example data has been randomly made. On average, the dialogue transcriptions consisted of 1209 words ($SD = 411$), ranging between 606 and 1869 words. The manually created SOAP reports consisted, on average, of 60 words ($SD = 17$), ranging between 37 and 87 words.

3.3 Analysis of Prompts

Despite growing interest in ML as a tool for medicine, there is a lack of knowledge about how these models operate and how to properly assess them using various metrics (Hicks et al., 2022). In this study, the resulting automated reports were evaluated against the human

reference reports using accuracy metrics and a human evaluation. By combining these quantitative and qualitative insights, this two-step review approach gives a comprehensive assessment of the automated reports' performance.

3.3.1 Accuracy Measurement

We used ROUGE as an accuracy metric since this is the most used text summarization evaluation metric to automatically evaluate the quality of a generated summary by comparing it to a human reference and it is suitable for our Dutch reports (Barbella and Tortora, 2022; Tangsali et al., 2022). The ROUGE metric code offered by the HuggingFace library was used to calculate the ROUGE1, and ROUGEL scores of the automated medical reports (Lin, 2004). ROUGE1 assessed the unigram similarities, and ROUGEL the longest common subsequence of words, between the automated report and the human reference reports (Tangsali et al., 2022).

The generation of the automated medical report is stochastic because generative AI models frequently display variety in their replies to a given prompt. To account for this variability every prompt was run five times on all transcripts, yielding distinct responses with each run. The prompts were run five times to strike a balance between robustness and computational efficiency, taking the trade-off between thorough analysis and computational costs into account. For every run, ROUGE was calculated. The overall performance and consistency of the automated medical reports are indicated by computing the average ROUGE score per consultation. Finally, an overall mean of these averages, with their standard deviations, was calculated and presented in the findings.

3.3.2 Human Evaluation

It is important to note that none of the automatic evaluation metrics are perfect and that human evaluation is still essential to ensure the quality of generated summaries (Falcão, 2023). For the human evaluation, the generated reports were manually analyzed. The words in the reports were categorized into three groups based on whether they were identical, paraphrased, or additional to the human reference. We also identified and classified the additional statements in the automatic reports into distinct categories. The identified categories were: duration of complaints, duration of treatment, previously tried treatments, doctor's observations, specific complaints (all reported symptoms by the patient), referral to which hospital, wait for results, discussed treatment (all specific steps that the GP reports to the patient),

expected patient actions, and other complaints that are ultimately not related to the diagnosis made in the human reference. Based on the clinical report idea of (Savkov et al., 2022) six medical professionals were asked to evaluate the importance of these classified additions in a SOAP report. Based on the response of the medical professionals, the additions were classified according to an adapted version of the taxonomy of error types by (Moramarco et al., 2022). Not all of their errors were observed in our study, besides, we replaced their “incorrect order of statements error” with a “categorization error”, and we identified “redundant” statements additionally.

4 FINDINGS AND DISCUSSION

Running of the formulated prompts, resulted in automated medical reports with wordcounts shown in Table 2. The automated reports are approximately twice as long as the human references, indicating a significant disparity in the length of the generated content. This could be explained by the fact that GPT generates full sentences, providing more detailed descriptions, while GPs tend to use abbreviations and keywords to convey the same information more concisely. Four out of six GPs in the expert panel indicated that they prefer abbreviations and keywords over full sentences, however, one GP preferred full sentences.

4.1 Accuracy Measurement

In the evaluation of the accuracy of the prompts, first, the *shot-prompting* technique was evaluated, followed by the *context manager pattern* technique that built on the optimal numbers of shots.

4.1.1 Shot-Prompting

Table 3 shows the comparison of the different shot-prompting approaches. The comparison shows that the *zero-shot prompting* approach resulted in the lowest ROUGE scores (0.121 and 0.079). *One-shot prompting* resulted in slightly higher scores (0.150

Table 2: Word count comparison between the generated report and the human reference.

	Human Reference	Generated Report	Difference
Subjective	29	47	18
Objective	11	20	9
Analysis	4	8	4
Plan	14	33	19
Total	111	58	53

Table 3: Mean and Standard Deviation (SD) for the ROUGE1 and ROUGE L-scores for zero-shot, one-shot, and two-shot prompting.

	ROUGE1 Mean±SD	ROUGEL Mean±SD
Zero-shot	0.121±0.007	0.079±0.006
One-shot	0.150±0.009	0.104±0.006
Two-shot	0.174±0.005	0.123±0.004

and 0.104) and the *two-shot prompting* approach resulted in the highest ROUGE scores (0.174 and 0.123). This result shows that adding shots to a prompt improves the performance. This can be explained by the fact that the shots serve as a reference for the expected output, enabling the GPT to generate similar outputs, which is in line with earlier research (Reynolds and McDonnell, 2021). Adding an increasing number of shots could result in higher performances than two-shots since few-shot prompting is generally meant to include a larger set of examples (Brown et al., 2020). This was, however, not possible due to the limited data set. Controversially, using fewer examples, makes it possible to create more well-crafted examples and comes closer to human performance (Brown et al., 2020). Additionally, (Zhao et al., 2021) found that few-shot prompting might introduce biases into certain answers.

It is also worth considering that the absence of a universally accepted method for applying shot prompting introduces a degree of uncertainty regarding the most effective approach. Including the transcripts with the sample SOAP reports, rather than only presenting the SOAP report as an example could have potentially produced different results. However, it is important to note that the main goal of this study was to teach the GPT how to correctly use the SOAP format and how to describe items in the SOAP categories.

4.1.2 Context Manager Pattern

The two-shot prompting strategy produced the highest scores, thus the *context manager pattern* was added to this foundation. Scope context and domain context were evaluated separately as well as the combination of the two types of context. In the assessment of the context manager pattern, a slight variation could be observed in the ROUGE scores based on different contextual additions (Table 4).

The combined *scope context* (0.179 and 0.126) scored lower than the combined *domain context* (0.220 and 0.167). This would suggest that *scope context* has little effect on the quality of reports that are generated. However, the combination of *scope context* and *domain context* (0.250 and 0.189) resulted in higher ROUGE scores than *domain context*

Table 4: Mean and Standard Deviation (SD) for the ROUGE1 and ROUGEL-scores for context prompts.

	ROUGE1 Mean±SD	ROUGEL Mean±SD
Context: Scope		
Context a	0.172±0.041	0.120±0.016
Context b	0.173±0.043	0.124±0.022
Context a & b	0.179±0.049	0.126±0.023
Context: Domain		
Context c	0.242±0.035	0.179±0.016
Context d	0.173±0.048	0.121±0.025
Context c & d	0.220±0.064	0.167±0.037
Context: Total		
Context a & b & c & d	0.250±0.049	0.189±0.025

by itself. A noteworthy finding is the difference between *domain contexts c* and *domain context d*, where *context d* produced lower scores (0.173 and 0.121) than *context c* (0.242 and 0.179). Remarkably, *contexts c* and *context d* together (0.220 and 0.167) also produced lower results than *context c* by itself. This suggests a potential negative effect of *context d* on the overall performance. To test this, a prompt was run that excluded *context d* from the prompt but this led to even lower overall scores (0.239 and 0.178). This decline in score may be explained by the limited dataset, which could have resulted in skewed results.

A potential reason why domain context increases the performance more than scope context is that the shot prompting already provides clear direction on how the GPT should behave; it has already set the context to the medical field. Prompting to use abbreviations, short sentences, and keywords (*context c*), may have had a considerable influence since GPT itself tends to make long sentences and provide as much information as possible. Prohibiting this action resulted in improved performance in the automated report. It is notable that it is unexpected that the GPT does not already do this after the shot prompting, but this could possibly be explained because only SOAP examples were used without including the transcripts in the examples.

This study also investigated the inclusion of a list of abbreviations within the prompt and found that it had a positive impact on the results, with ROUGE scores of 0.273 and 0.261. However, it was ultimately not selected as the optimal prompt since the use of abbreviations varies between hospitals and healthcare providers (Borcherding and Morreale, 2007), making it difficult to create a universally applicable prompt that incorporates all relevant abbreviations.

4.2 Human Evaluation

The results from the quantitative approach showed that the *two-shot prompting* approach in combination

with the *scope* and *domain context* (Listing 5) resulted in the best performance. However, since this still resulted in a relatively low ROUGE score, human evaluation was performed for this final prompt.

- 1 Within the scope of medical dialogue summarization, create a medical report from the consultation, follow the SOAP guidelines, and use the information found in the provided transcript as the sole source.
- 2 Consider that you are a general practitioner who writes the medical report.
- 3 Consider that the report is used for communication between doctors who use abbreviations and short sentences.
- 4 Consider that in the medical field, the division between left and right, and the medication dosage are important.
- 5 Use the following examples: [example1], [example2].

Listing 5: The best performing prompt.

The expert panel showed that all six GPs agreed on the fact that the duration of the complaints is relevant to mention within the report. For all the other categories there seems to be disagreement about the relevance. For example, there appears to be disagreement about the importance of recording specific patient complaints. When mentioning that in particular, the left ear caused problems, the GPs disagree on the importance. Some indicate that this is relevant ($n = 3$), while there are also GPs that indicate that

Human Reference

S: Reduced hearing and sensitivity le/ri. Nose drops and syringing no effect.
 O: AD : redness and swelling ear canal and redness and flaking auricle. AS : redness auricle.
 E: Otitis externa both sides
 P: Sofradex 2 dd 2 drops. apply cream to the skin, revision 1 week

Generated Report

S: Patient states deafness and sensitivity in the ear since syringing by assistant. No severe pain. Patient has tried nose drops on their own without improvement.
 O: Ear canal infected, right worse than left. Flaking in the auricle. Eardrum visible, but auricle narrow.
 E: External ear canal infection, bilateral, right worse than left.
 P: Prescribed Sofradex ear drops, 2dd2 drops. Also cream for dry auricle. Patient asked to come back next week for control.

Legend:

Equivalent	Addition	Error		
Identical	Relevant	Hallucination	Categorization	Repetition
Paraphrase	Redundant	Incorrect	Omission	

Figure 3: Human evaluation of the automated medical report of transcript 2028 (translated to English, the generated reports are in Dutch).

Table 5: Error statements with occurrences in the five generated medical reports (translated to English, the generated reports are in Dutch).

Type	Definition - Examples	Occurrence
Factual Errors	An error in the information presented that contradicts reality.	14
Hallucinations	<i>"Pain originating from the syringing by the doctor's assistant"</i> Pain was already present before the syringing.	6
Incorrect statements	<i>"The patient uses Rhinocort and cetirizine daily for mucous membrane hyperreactivity"</i> Patient only uses Rhinocort for mucous membrane hyperactivity.	8
Stylistic Errors	An error in the manner in which information is used or presented.	17
Repetitions	<i>"Patient feels sick"</i> <i>"Patient also reports a feeling of being unwell"</i> .	3
Classification error	<i>"The area around the ear feels numb."</i> in the Analysis part of SOAP.	14
Omissions	An error characterized by the act of neglecting to include essential information in the report.	19
In Subjective	Indication of which ear is involved/ referred to Parts of symptoms mentioned Parts of relevant medical history	3 2 5
In Objective	Indication of which ear is involved/ referred to Parts of symptoms observed	2 2
In Analysis	Indication of which ear is involved/ referred to	3
In Plan	Agreement with patient Possible future treatment	1 1
Redundant Statements	The inclusion of unnecessary information that does not contribute substantively to the report, although it is on the topic of the medical condition.	25
In Subjective	<i>The patient reports ... especially in the morning, and that the ear smells.</i>	7
In Objective	<i>Left: some earwax.</i>	5
In Analysis	<i>This can also radiate from the sinuses.</i>	2
In Plan	<i>A dressing and plaster have been applied to the left ear to collect the discharge.</i>	9
Additional	<i>Colonoscopy scheduled for three years. Patient should contact for referral to a gastroenterologist. Prescription for [name of medication] for constipation.</i> In an additional NB (Nota Bene)	2

The occurrence is counted per consultation, so if the same error happened repeatedly in the reruns for the same consultation, it was only counted once.

this is not relevant ($n = 1$) or they indicate that they are neutral about this ($n = 2$). However, one of the GPs who indicated that it is relevant did mention that they would note it more briefly. Another example that shows this disagreement is within the discussed treatment: "gauze and plaster applied to the left ear to collect discharge". Two GPs indicated that this was relevant, two indicated that this was irrelevant and two indicated that they were neutral about this.

Table 5 shows the identified error statements in the five automated reports during the human evaluation. The human evaluation highlights several noteworthy findings regarding the quality of the automated reports. It is evident that the automated reports contain a notable number of redundant statements, 25 in total, with the majority occurring in the **Plan** section ($n = 9$) and the **Subjective** section ($n = 7$). Moreover, stylistic errors are prevalent, particularly classification errors ($n = 14$) and occasional repetitions ($n = 3$). In addition to adding extra (relevant or redundant) informa-

tion, the automated report sometimes omits essential information ($n = 19$) when compared to the human reference. Factual errors are present as well, amounting to a total of 8 incorrect statements and 6 hallucinations. For a visual example of the error statements see Figures 3 and 4.

A possible reason for the omissions in the automated reports could be related to the GPT's limited understanding of the medical context, leading it to overlook certain critical details during the report generation process. This is supported by research from (Johnson et al., 2023), who found that a potential limitation of GPT is handling complex medical queries, but they did not reach statistical significance for this statement. A potential reason for the classification errors is that the GPT lacks genuine comprehension of the distinct SOAP categories, thus negatively influencing its ability to accurately allocate information to the appropriate category within the SOAP report.

The human evaluation revealed substantial vari-

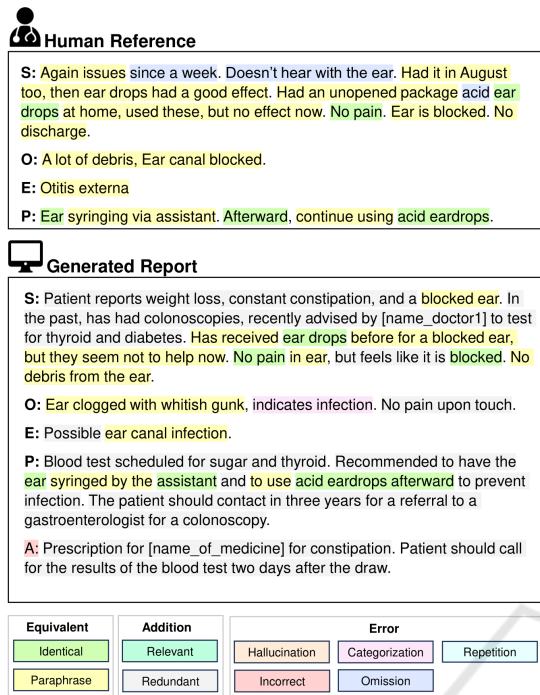


Figure 4: Human evaluation of the automated medical report of transcript 2006 (translated to English, the generated reports are in Dutch).

ations in the performance of the automated reports across different consultations, with some reports displaying higher performance levels than others. For example, the report that was generated based on transcript 2006 (Figure 4) had a lot of redundant information added to the report while the report based on transcript 2808 closely resembles the human references (Figure 3). This discrepancy in performance may be related to the difficulty that GPT encounters in differentiating between various medical conditions discussed during a single consultation, which may result in the creation of SOAP reports that include data pertaining to several medical conditions.

One noteworthy finding is that, even though the prompts make clear how important left-right orientation is, the automated reports frequently miss it. This can be explained by the findings from the research of (Ye and Durrett, 2022); in their research, they have demonstrated that adding explanations or contextual cues alone does not necessarily ensure an improved result in the final output. This underlines the problem of ensuring that complicated, contextually relevant information is consistently included in the generated reports. This disparity highlights the ongoing difficulties in optimizing automated report production for medical contexts and argues the natural language processing system’s flexibility and understanding when it comes to adding important facts.

5 CONCLUSION

Even though machine learning is becoming more popular as a medical tool, little is known about these models’ workings or how to appropriately evaluate them using different metrics. In this research, we investigated the combination of shot-prompting with pattern prompting to enhance the performance of automated medical reporting. The automated medical reports were generated with the use of prompt engineering software. The generated reports were evaluated against human reference provided by a GP. For this evaluation, the widespread ROUGE metric was used in combination with human evaluations. The results showed that adding examples to the prompt is beneficial for the automated medical report. It also showed that adding both scope context as well as domain context improved the performance of the automated medical report. This resulted in the overall best structure for a prompt using a base structure in combination with two shots and scope and domain context.

5.1 Limitations

Despite these promising results, this study has validity threats that could have influenced the findings. Firstly, generative AI systems are stochastic which introduces variability as they produce different answers each time they are run, which may impact the reliability and repeatability of the results. Secondly, the findings have limited generalizability to other medical conditions because of the constrained data availability, with a small dataset exclusively on Otitis, and the variability in medical reporting across diverse domains. An additional concern is the missed opportunity to explore every combination of shots and contexts. However, the feasibility of this approach was constrained within the scope of this study. This influenced the study’s depth of analysis and its capacity to provide nuanced insights. Lastly, the human evaluation has some limitations, even though medical professionals were consulted to gather domain expertise the human evaluation was still performed by non-medical professionals. This potentially introduced a perspective misalignment that could have influenced the interpretation and assessment of the generated medical reports.

5.2 Future Work

This marks an initial investigation into optimizing prompt sequences with a fixed LLM. Nonetheless, we acknowledge that diverse LLMs may yield different outcomes. Additionally, future studies should explore

the applicability of our findings in the setting of different medical conditions and broaden the scope of the study beyond Otitis. The prompt could be further improved to avoid redundant statements by defining the maximum length of the output, using an increasing number of shots, or using a different method of shots such as providing the consultation transcript in addition to the resulting medical report.

Furthermore, future work should focus on finding a more suitable metric to evaluate the output. In the current research, the ROUGE metric was used for the evaluation of the automated medical report as well as human evaluation. ROUGE is commonly used within summarization tasks however it has some downsides, the metric is very black and white. It does not take into account the meaning of the words in the summarization but only the occurrence of specific words. For future work a different evaluation needs to be created, this metric needs to take into account the meaning of the automated medical report, and it needs to investigate if the essence of the automated medical report matches the golden standard. This new metric needs to take into account rewording and paraphrasing so that they are not automatically considered wrong. For optimal evaluation, the complete reports should be evaluated by GPs.

ACKNOWLEDGEMENTS

We want to thank all the GPs and other medical professionals who aided us in our human evaluation. Special thanks go to Rob Vermond for assisting with the expert panel of the GPs. Their professional insights ensured that we could execute the human evaluation. We also would like to thank Kate Labunets for providing feedback on the paper. Finally, many thanks go to Bakkenist for the support of this research project.

REFERENCES

- Anil (2023). Prompt engineering -1- shot prompting. *Medium*. <https://tinyurl.com/shotprompting>.
- Balagurunathan, Y., Mitchell, R., and El Naqa, I. (2021). Requirements and reliability of ai in the medical context. *Physica Medica*, 83:72–78.
- Banerjee, D., Singh, P., Avadhanam, A., and Srivastava, S. (2023). Benchmarking llm powered chatbots: Methods and metrics. *arXiv preprint arXiv:2308.04624*.
- Barbella, M. and Tortora, G. (2022). Rouge metric evaluation for text summarization techniques. *Available at SSRN 4120317*.
- Bigelow, S. J. (2023). 10 prompt engineering tips and best practices: Techtargget. <https://tinyurl.com/prompt-bestpractices>.
- Borcherding, S. and Morreale, M. J. (2007). *The OTA's guide to writing SOAP notes*. Slack Incorporated.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Coorevits, P., Sundgren, M., Klein, G. O., Bahr, A., Claerhout, B., Daniel, C., Dugas, M., Dupont, D., Schmidt, A., Singleton, P., et al. (2013). Electronic health records: new opportunities for clinical research. *Journal of internal medicine*, 274(6):547–560.
- Dong, Q., Li, L., Dai, D., Zheng, C., Wu, Z., Chang, B., Sun, X., Xu, J., and Sui, Z. (2022). A survey for in-context learning. *arXiv preprint arXiv:2301.00234*.
- Dragon, D. (2023). The right way to do few-shot prompting. *Medium*. <https://tinyurl.com/few-shot-prompting>.
- ElAssy, O., de Vendt, R., Dalpiaz, F., and Brinkkemper, S. (2022). A semi-automated method for domain-specific ontology creation from medical guidelines. In *International Conference on Business Process Modeling, Development and Support*, pages 295–309. Springer.
- Falcão, F. (2023). Metrics for evaluating summarization of texts performed by transformers: how to evaluate the quality of summaries. <https://tinyurl.com/metrics-quality>.
- Friedberg, M. W., Chen, P. G., Van Busum, K. R., Aunon, F., Pham, C., Caloyeras, J., Mattke, S., Pitchforth, E., Quigley, D. D., Brook, R. H., et al. (2014). Factors affecting physician professional satisfaction and their implications for patient care, health systems, and health policy. *Rand health quarterly*, 3(4).
- Heston, T. F. (2023). Prompt engineering for students of medicine and their teachers. *arXiv preprint arXiv:2308.11628*.
- Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., and Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. *Scientific reports*, 12(1):5979.
- Houwen, J., Lucassen, P. L., Stappers, H. W., Assendelft, W. J., van Dulmen, S., and Olde Hartman, T. C. (2017). Improving gp communication in consultations on medically unexplained symptoms: a qualitative interview study with patients in primary care. *British Journal of General Practice*, 67(663):716–723.
- Jain, R., Jangra, A., Saha, S., and Jatowt, A. (2022). A survey on medical document summarization. *arXiv preprint arXiv:2212.01669*.
- Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., and Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.
- Johnson, D., Goodman, R., Patrinely, J., Stone, C., Zimmerman, E., Donald, R., Chang, S., Berkowitz, S., Finn, A., Jahangir, E., et al. (2023). Assessing the accuracy and reliability of ai-generated medical responses: an evaluation of the chat-gpt model. *Research square*.

- Krishna, K., Khosla, S., Bigham, J. P., and Lipton, Z. C. (2021). Generating soap notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 4958–4972.
- Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., and Neubig, G. (2023a). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.
- Liu, Y., Iyer, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. (2023b). Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.
- Maas, L., Geurtsen, M., Nouwt, F., Schouten, S., Van De Water, R., Van Dulmen, S., Dalpiaz, F., Van Deemter, K., and Brinkkemper, S. (2020). The care2report system: Automated medical reporting as an integrated solution to reduce administrative burden in healthcare. In *HICSS*, pages 1–10.
- Mathur, Y., Rangrejji, S., Kapoor, R., Palavalli, M., Bertsch, A., and Gormley, M. (2023). Summqa at medqa-chat 2023: In-context learning with gpt-4 for medical summarization. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.
- Meijers, M. C., Noordman, J., Spreeuwenberg, P., Olde Hartman, T. C., and van Dulmen, S. (2019). Shared decision-making in general practice: an observational study comparing 2007 with 2015. *Family practice*, 36(3):357–364.
- Michalopoulos, G., Williams, K., Singh, G., and Lin, T. (2022). Medicalsum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4741–4749.
- Mickey, N. (2023). Explore the benefits of azure openai service with microsoft learn: Azure blog: Microsoft azure. *Azure*. <https://tinyurl.com/azure-openai>.
- Moramarco, F., Korfiatis, A. P., Perera, M., Juric, D., Flann, J., Reiter, E., Savkov, A., and Belz, A. (2022). Human evaluation and correlation with automatic metrics in consultation note generation. In *ACL 2022: 60th Annual Meeting of the Association for Computational Linguistics*, pages 5739–5754. Association for Computational Linguistics.
- Mridha, M. F., Lima, A. A., Nur, K., Das, S. C., Hasan, M., and Kabir, M. M. (2021). A survey of automatic text summarization: Progress, process and challenges. *IEEE Access*, 9:156043–156070.
- Nair, V., Schumacher, E., and Kannan, A. (2023). Generating medically-accurate summaries of patient-provider dialogue: A multi-stage approach using large language models. *arXiv preprint arXiv:2305.05982*.
- Overhage, J. M. and McCallie Jr, D. (2020). Physician time spent using the electronic health record during outpatient encounters: a descriptive study. *Annals of internal medicine*, 172(3):169–174.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.
- Reynolds, L. and 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*, pages 1–7.
- Robinson, R. (2023). How to write an effective gpt-3 or gpt-4 prompt. *Zapier*. <https://tinyurl.com/gpt-prompt>.
- Savkov, A., Moramarco, F., Korfiatis, A. P., Perera, M., Belz, A., and Reiter, E. (2022). Consultation checklists: Standardising the human evaluation of medical note generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 111–120.
- Schenker, J. D. and Rumrill Jr, P. D. (2004). Causal-comparative research designs. *Journal of vocational rehabilitation*, 21(3):117–121.
- Tam, A. (2023). What are zero-shot prompting and few-shot prompting. *Machine Learning Mastery*. <https://tinyurl.com/machine-learning-mastery>.
- Tangsal, R., Vyawahare, A. J., Mandke, A. V., Litake, O. R., and Kadam, D. D. (2022). Abstractive approaches to multidocument summarization of medical literature reviews. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 199–203.
- van Buchem, M. M., Boosman, H., Bauer, M. P., Kant, I. M., Cammel, S. A., and Steyerberg, E. W. (2021). The digital scribe in clinical practice: a scoping review and research agenda. *NPJ digital medicine*, 4(1):57.
- Wang, J., Shi, E., Yu, S., Wu, Z., Ma, C., Dai, H., Yang, Q., Kang, Y., Wu, J., Hu, H., et al. (2023). Prompt engineering for healthcare: Methodologies and applications. *arXiv preprint arXiv:2304.14670*.
- White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., and Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.
- Widyassari, A. P., Affandy, A., Noersasongko, E., Fanani, A. Z., Syukur, A., and Basuki, R. S. (2019). Literature review of automatic text summarization: research trend, dataset and method. In *2019 International Conference on Information and Communications Technology (ICOIACT)*, pages 491–496. IEEE.
- Yadav, D., Desai, J., and Yadav, A. K. (2022). Automatic text summarization methods: A comprehensive review. *arXiv preprint arXiv:2204.01849*.
- Ye, X. and Durrett, G. (2022). The unreliability of explanations in few-shot prompting for textual reasoning. *Advances in neural information processing systems*, 35:30378–30392.
- Zhao, Z., Wallace, E., Feng, S., Klein, D., and Singh, S. (2021). Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR.