MaLei at MultiClinSUM: Summarisation of Clinical Documents using Perspective-Aware Iterative Self-Prompting with LLMs

Notebook for the <MultiClinSUM> Lab at CLEF 2025

Libo Ren[1], Yee Man Ng[2] and Lifeng Han[1,2,3,*]

[1]University of Manchester, UK
[2]Leiden Institute of Advanced Computer Science (LIACS), Leiden University, The Netherlands
[3]Leiden University Medical Center, The Netherlands

Abstract

Efficient communication between patients and clinicians plays an important role in shared decision-making. However, clinical reports are often lengthy and filled with clinical jargon, making it difficult for domain experts to identify important aspects in the document efficiently. This paper presents the methodology we applied in the MultiClinSUM shared task for summarising clinical case documents. We used an Iterative Self-Prompting technique on large language models (LLMs) by asking LLMs to generate task-specific prompts and refine them via example-based few-shot learning. Furthermore, we used lexical and embedding space metrics, ROUGE and BERT-score, to guide the model fine-tuning with epochs. Our submission using perspective-aware ISP on GPT-4 and GPT-4o achieved ROUGE scores (46.53, 24.68, 30.77) and BERTscores (87.84, 83.25, 85.46) for (P, R, F1) from the official evaluation on 3,396 clinical case reports from various specialties extracted from open journals. The high BERTscore indicates that the model produced semantically equivalent output summaries compared to the references, even though the overlap at the exact lexicon level is lower, as reflected in the lower ROUGE scores. This work sheds some light on how perspective-aware ISP (PA-ISP) can be deployed for clinical report summarisation and support better communication between patients and clinicians.

Keywords

Shared Decision Making, Health Literacy, Patient Communication, LLMs, Clinical Summarisation

CEUR-WS.org/Vol-4038/paper_37.pdf

1. Introduction

Efficient and effective communications between patients and healthcare professionals play an important role in patient care [1, 2]. However, healthcare providers frequently have to read many clinical documents in a short time frame to understand the current patients. This is challenging, as clinical documents of patients include rich information on patients' problems, diagnoses, treatments, progressions, and side effects. Similarly, patients often do not have the clinical expertise to fully understand the lengthy clinical documents about their health issues. A concise and accurate summarisation of clinical documents will save the time of healthcare professionals to understand the problem at hand, as well as help patients to understand their health conditions better and earlier. We attended the multilingual clinical documents summarisation shared task (MultiClinSUM) to explore large language models (LLMs) for the use of this challenge.

The MultiClinSUM Track is organised by the Barcelona Supercomputing Center's NLP for Biomedical Information Analysis group and promoted by Spanish and European projects such as DataTools4Heart, AI4HF, and BARITONE [3].

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

[*] Corresponding author.

renlibo994@gmail.com (L. Ren); y.m.ng@liacs.leidenuniv.nl (Y. M. Ng); l.han@lumc.nl (L. Han)
https://github.com/Libo-Ren (L. Ren); https://www.universiteitleiden.nl/en/staffmembers/yee-man-ng (Y. M. Ng); https://www.universiteitleiden.nl/en/staffmembers/lifeng-han (L. Han)
0009-0005-9182-9424 (L. Ren); 0009-0008-9659-6510 (Y. M. Ng); 0000-0002-3221-2185 (L. Han)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

To utilize the current state-of-the-art development from natural language processing (NLP), we investigated Iterative Self-Prompting (ISP) [4] in GPT-4 and ChatGPT for automatic summary of clinical documents. In this methodology, we ask the LLMs to generate prompts themselves to approach this task by detailed instruction and example-based learning; the generated automatic summary is returned to the LLMs with example references to ask the LLM to refine the prompts for this specific task. The ISP technique has proved to be very useful and efficient in leveraging LLMs to generate a clear summary that includes patients' symptoms, diagnoses, treatments, and outcomes/follow-ups, in a previous shared task for healthcare answer summarization [4]. We use the datasets provided by the shared task organizers, which contain clinical case reports and summaries that are written by healthcare providers and in various languages, including English, Spanish, French, and Portuguese. The English test set under investigation contains 3,396 clinical documents. We perform both quantitative and qualitative evaluations, through human and automatic metrics, including ROUGE and BERTscore. ROUGE is a lexical overlapping metric, and BERTscore is an embedding space semantic similarity metric.

Quantitative evaluation scores show that while our system output has a ROUGE F1 score of 0.31 compared to the reference on exact lexicon matches, the semantic BERTscore shows 0.85 F1, which indicates high quality of semantic meaning preservation. This sheds good light on the potential usage of LLMs for the summarisation of clinical documents using the ISP technique. The qualitative analysis confirms that the generated summaries tend to cover the key clinical aspects and contain logical paraphrasing. We also carried out error analysis to see in what ways LLMs produce undesirable results, such as the considerable length of the summaries. Another interesting finding from this shared task is that LLMs tend to generate longer text to comment on missing data when the clinical document is too short, such as "The case report does not provide specific details on the outcome or follow-up. Typically, such a patient would require close monitoring and treatment adjustments based on laboratory and clinical responses." This also provides some insight into the clarity of current clinical documents/reports.

2. Background and Related Work

2.1. Clinical NLP

Clinical NLP has drawn attention from both NLP and healthcare researchers in recent years due to the development and effectiveness of modern NLP models and the eagerness to test such AI models in healthcare domains. For instance, there is the ClinicalNLP workshop series from 2016 (6th edition in 2024) [5]. A related track of ClinicalNLP WS is the Biomedical NLP international workshop events (BioNLP) from 2004 [6, 7].

The corresponding tasks have included clinical text translation [8], biomedical abstract simplification [9, 10], clinical events recognition [11], temporal relation extraction [12, 13], entity linking/normalisation to SNOMED CT and British National Formulary (BNF) codes [14], synthetic data generation/augmentation [15, 16], patient sensitive information de-identification [17], and healthcare answer summarisation [4], etc. These works have explored methodologies from different paradigms, such as fine-tuning encoder-based models, training encoder-decoder models, and prompting decoder-only models using different techniques.

2.2. Healthcare Data Summarisation

For the clinical documents summarisation task, the most relevant work includes the shared task on Perspective-Aware Healthcare Answer Summarisation (PerAnsSumm 2025) [18]. This shared task included the summarisation of online forum healthcare answers while considering the different perspectives, i.e., types of information such as 'Cause" or 'Suggestion', within an answer. In this shared task, we used Iterative Self-Prompting (ISP) with Claude and o1 for Perspective-aware Healthcare Answer Summarisation, as described in [4, 18]. Similar to clinical documents, online forum responses vary greatly in length. The key difference is that clinical documents often contain clinician-specific abbreviations and jargon, which pose challenges for NLP models to interpret. In contrast, online forum data typically

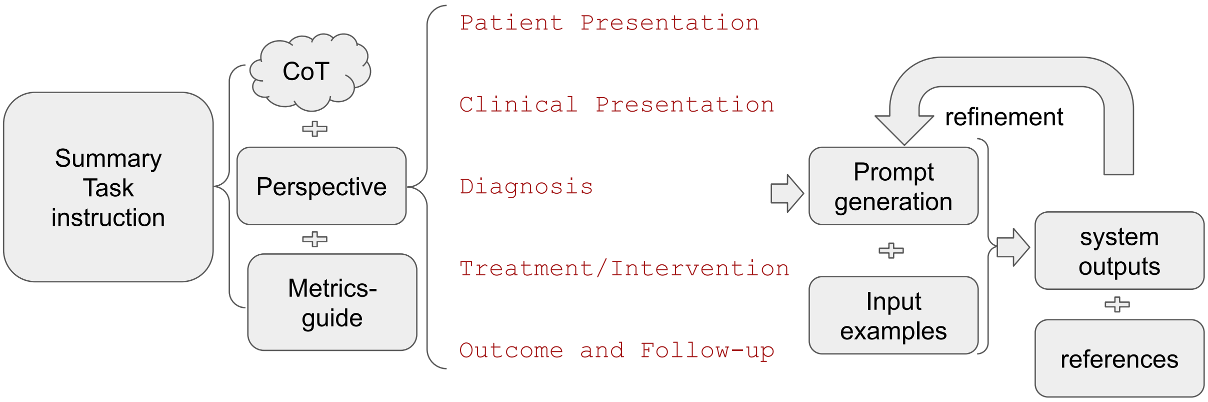


Figure 1: Perspective-aware Iterative Self-Prompting (PA-ISP) Illustration Diagrams from the MaLei team.

includes social media-style writing, such as frequent spelling mistakes and grammatical errors. In this work, we build upon our experience from the PerAnsSumm shared task to design perspective-aware summaries for clinical documents.

3. Methodology

3.1. Overview of the Prompting Framework

There are mainly three clinical datasets involved in this project, as listed below:

- *multiclinsum_gs_train_en:* contains 592 gold-standard samples, which are manually annotated and consist of a full-text and summary pair.
- *multiclinsum_large-scale_train_en:* contains 25,902 full-text and summary pairs. Their quality is slightly lower than that of the 592 gold samples, but they are still useful for data augmentation.
- *multiclinsum_test_en:* the English test set, which includes 3,396 full-text cases but without any summaries.

We mainly adopt the Iterative Self-Prompting (ISP) strategy in this task. As shown in Figure 1, we construct the meta-prompt based on the combination of Chain-of-Thought (CoT) instructions, clinical perspectives, and metric-based guidance. The meta-prompt is provided to the LLM along with a few few-shot examples at the beginning.

Based on this meta-prompt and the examples, the LLM is instructed to generate a new task-specific prompt that guides the clinical summarisation process more effectively. Using this synthetic prompt (prompt_v1), we input it together with clinical full-texts from a portion of the golden training set into the model to generate corresponding summaries. These synthetic summaries are then compared with the ground-truth summaries from the golden data, and evaluation scores—as well as reflective feedback and suggestion advice—are produced accordingly. This feedback serves as a reference for further prompt refinement, allowing us to iteratively update the prompt to obtain prompt_v2, _v3, and so on.

This prompt updating process is repeated until no obvious performance improvement is observed. Once the improvement plateaued, we planned to augment the structure using additional spans extracted from the remaining gold data and to apply a retrieval-based (RAG) technique using the 25,902 non-golden training samples. After all experiments, the best-performing prompt version is used to generate the final clinical summaries on the test set.

Unfortunately, due to time constraints, we were only able to complete the steps before the structure augmentation phase (as detailed in Section 3.4). As a result, we selected the best-performing prompt at that point and used it directly for inference on the test set. The remaining experimental designs can be explored in future work.

3.2. Prompt Initialisation and Few-shot Setup

As shown in Figure 1, we first construct an initial instruction that prompts the LLMs about the summary task description with the chain of thoughts (CoTs) on how it shall think, for example:

- What common structure or patterns do you observe in the examples?
- What information is emphasised?
- How can a language model be guided to produce similar quality outputs?
- What errors should be avoided?

These CoTs are combined with perspective-based, i.e., multifaceted, structural guidance and metric-based feedback to inform the LLM's generation process.

The Perspectives we designed include:

1. Patient Presentation: age, sex, relevant history.
2. Clinical Presentation: key symptoms and signs.
3. Diagnosis: relevant investigations, tests, conclusions.
4. Treatment/Intervention: medications, surgeries, therapies.
5. Outcome and Follow-up: results of treatment, current status.

The metrics we used are ROUGE-L (for lexical overlap) and BERTScore (for embedding semantic fidelity). The instructions, structural perspectives, evaluation metrics, and three representative examples collectively form the meta-prompt, which is then used to generate the initial prompt. An example of the meta-prompt is provided in Appendix B.1.

3.3. Prompt Iterative Update and Refinement

We selected a small batch of 50 full-text and summary pairs to train the prompt-interactive model. Note that data points indexed from 4 to 53 were used at this stage, as the first three samples had already been used for the initial prompt generation as few-shot examples.

In each epoch, we not only generated summaries based on the full texts using the initial prompt but also compared the synthetic summaries with the gold-standard annotations, asking GPT-4o to provide reflections and revision suggestions. For each generated summary, we computed both ROUGE-L and BERTScore, and requested reflections and revision suggestions from the model behind their performance. We found that BERTScore remained relatively stable (consistently above 0.85), while ROUGE-L scores fluctuated significantly, ranging from 0.12 to 0.52. Therefore, our optimisation efforts focused on improving ROUGE-L.

To iteratively update the prompt while balancing performance and computational cost, we selected a small subset of 15 summaries with the lowest ROUGE-L scores and included their corresponding evaluation feedback. These were used as new few-shot examples to guide the prompt refinement. The LLM was then instructed to revise the prompt by integrating the previous version along with the reflection and suggestion content of these samples. The prompt used in this process is shown in the Appendix B.2 Figure 9.

We conducted five epochs of this process. For the initial version (prompt_v1), one summary was found to be invalid for evaluation scoring, with an overall BERTScore of 0.86 and a ROUGE-L of 0.30 among the remaining 49 full-text and summary pairs. Interestingly, after the prompt updates, although the evaluation scores did not improve significantly, the invalid case was resolved. In other words, the BERTScore continued to fluctuate around 0.86, and the ROUGE-L around 0.30, across all 50 full-text and summary pairs. As a result, we adopted prompt_v2 as the best-performing version—the first in which invalid predictions were eliminated.

3.4. Full-data Structure-aware Prompt Enhancement

To further improve structural consistency, the remaining 539 gold summaries were intended for extracting common phrases and analysing section-wise linguistic patterns. For example, we aimed to identify common clinical spans and examine the language style of the gold-standard summaries, e.g., phrases like *"The patient presented with…"* or *"Treatments include…"*. We also considered measuring the average length of each paragraph and investigating whether consistent structural patterns could be observed and used to refine the prompt.

Based on these insights, the native instructions for the five clinical perspectives could be further specified. In addition, regular expressions or phrase-matching methods could be employed to capture high-frequency sentence structures or templates. Sentence patterns that are often overlooked could also be statistically analysed. Combining these steps may contribute to improving evaluation scores, particularly the ROUGE-L score.

Despite several iterations of prompt revision based on reflective feedback, the model outputs did not demonstrate any improvements in ROUGE-L, indicating that more sophisticated strategies beyond self-iterative prompt refinement may be required. Owing to time constraints, further exploration in this area is reserved for future work.

3.5. Similar Case Retrieval-based Few-shot Augmentation

In addition to the gold-standard set, the dataset also includes 51,804 extended clinical cases, each consisting of a full-text input and its corresponding summary. To enhance test-time generation, we retrieve cases whose input texts are semantically similar to the current full-test input using sentence embeddings. SentenceBERT, combined with cosine similarity, is applied at this stage.

The summaries from the top retrieved cases are inserted before the test input as few-shot demonstrations, following the same format as the manually selected gold examples at stage 3.2. Typically, the top 3 cases are selected to balance diversity with prompt length constraints. The retrieved summaries are not included in the evaluation and serve solely as auxiliary input for generation guidance.

3.6. Testset Inference

Once the optimal prompt was selected, the final version was used for testset summarization. The generation process follows the same setup described in Section 3.2. The main difference is that we expect the generated summary to be shorter than the original full text. Therefore, we compared the character lengths of the synthetic summaries and their corresponding full texts and identified cases where the summary was unexpectedly longer.

For these cases, we asked the language model to regenerate the summary up to five times. Some outputs were successfully shortened, while others remained longer than the input. In such cases, we directly replaced the generated summary with the original full text, assuming that the original text was already sufficiently concise.

It should be noted that although the full experimental pipeline was initially designed, we did not proceed with the Full-data Structure-aware Prompt Enhancement and Similar Case Retrieval-based Few-shot Augmentation (as detailed in Section 3.4 and Section 3.5) due to time constraints. These steps are planned for future work to further explore their potential for improving performance.

4. Experimental Work and Submission to MultiClinSUM

4.1. Development of the ISP-GPT-4/o model

For LLMs, we used GPT-4 to generate the initial prompt based on a meta-prompt, and GPT-4o for summarisation and reflection generation. We split the official MultiClinSUM data into several parts for training around 5 epochs, and selected a well-performing prompt for final test set inference, as described in Section 3. Figure 2 illustrates the data partitioning strategy and its corresponding usage

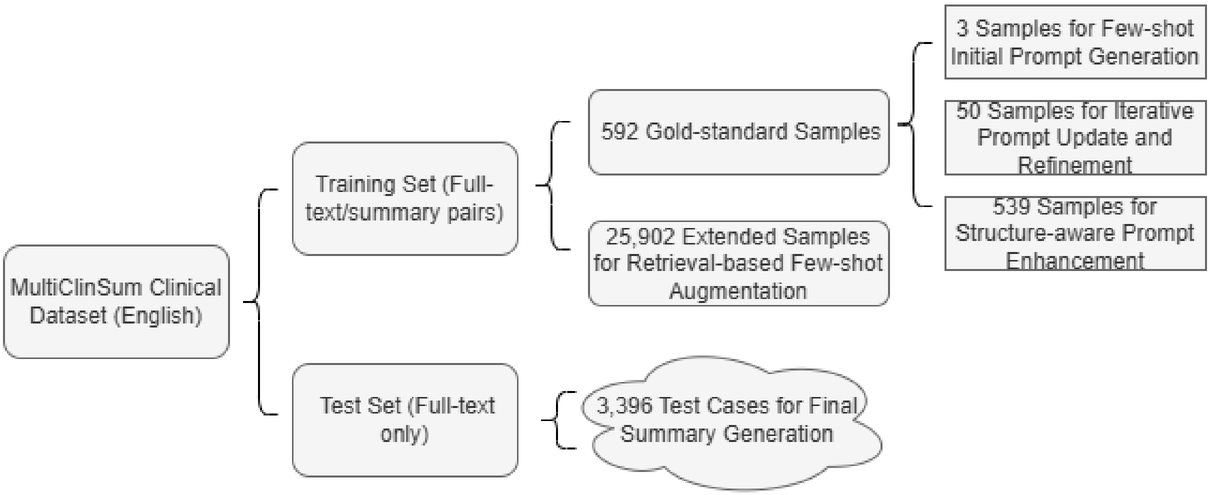

3 Samples for Few-shot Initial Prompt Generation
50 Samples for Iterative Prompt Update and Refinement
539 Samples for Structure-aware Prompt Enhancement
592 Gold-standard Samples
25,902 Extended Samples for Retrieval-based Few-shot Augmentation
Training Set (Full-text/summary pairs)
MultiClinSum Clinical Dataset (English)
Test Set (Full-text only)
3,396 Test Cases for Final Summary Generation

Figure 2: Data Partitioning and Usage in the MultiClinSUM Pipeline

in our pipeline. We completed summary generation for 3,396 English full-text cases. Other languages (Spanish, Portuguese, and French) will be explored in future work.

4.2. Submission outcome of MaLei from the shared task

4.2.1. Quantitative Results

For the MultiClinSUM2025 shared task we attended [3, 19], the results of 3,396 submitted English test summaries are shown in Table 1, and their corresponding Grouped Bar Chart and Overlaid Histogram are shown in Figure 3 and Figure 4.

Table 1

Evaluation metrics for BERTScore and ROUGE from MaLei team submissions.

Metric	Precision	Recall	F1 Score
BERTScore	0.8784	0.8325	0.8546
ROUGE-L	0.4653	0.2468	0.3077

At the test set level, as shown in Table 1 and Figure 3, BERTScore is overall more than twice as high as ROUGE-L, reflecting a similar trend observed in the training set. This suggests that our system achieves strong semantic preservation while tending to paraphrase the original full text using different linguistic styles. Another notable pattern further supports this. Across both metrics, precision consistently exceeds recall. This indicates that the synthetic summaries are generally accurate in terms of what they include, but may lack completeness at a finer-grained level. We speculate that this may be because GPT-4o tends to generate more concise or compressed text. Additionally, the ROUGE-L recall is particularly low — falling below 0.25 — which implies that the model often uses more varied expressions instead of preserving the original key phrases, leading to reduced lexical overlap. Therefore, future work could focus on identifying and preserving fixed phrases and structural patterns in the summary generation process. In summary, this result aligns with what has been observed in autoregressive models: they tend to focus more on what to generate rather than on precise word-by-word reproduction.

At the instance level, as shown in Figure 4, nearly all samples (3374 out of 3396, $\approx 99.35\%$) have BERTScores concentrated in the 0.8–0.9 range, indicating a high degree of semantic consistency. In contrast, most ROUGE-L scores fall between 0.2 and 0.4. The distribution exhibits a clear left-skew and a long-tail pattern, suggesting that ISP may have inaccurately generalized the structure of the examples provided to the model, thereby misleading the LLM's generation. Notably, 380 samples (11.19%) received ROUGE-L scores below 0.2, while none scored particularly low on BERTScore. This indicates that a subset of summaries may suffer from issues such as missing critical information, disorganized structure, or fragmented language.

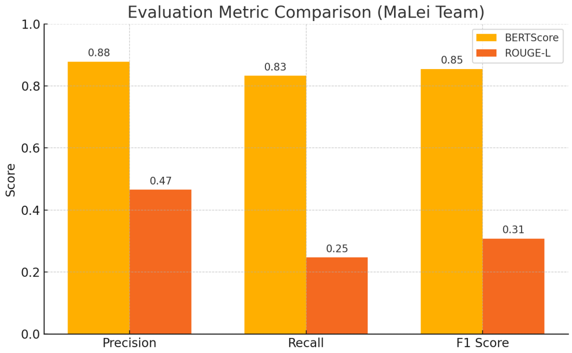


Figure 3: Evaluation Metric Comparison: BERTScore vs ROUGE-L

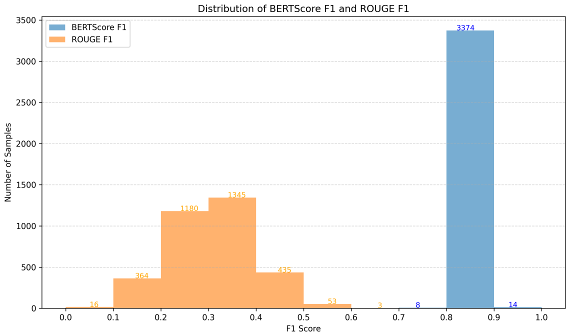
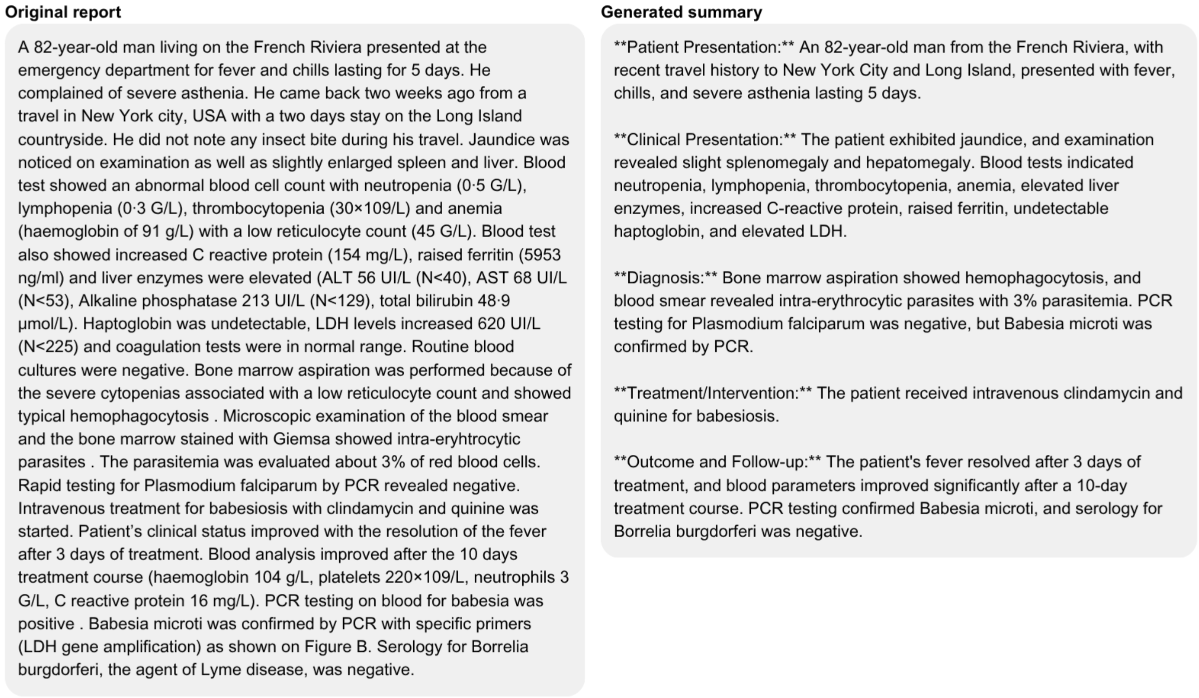


Figure 4: F1 Score Distribution: BERTScore vs ROUGE

4.2.2. Qualitative Results



Original report

A 82-year-old man living on the French Riviera presented at the emergency department for fever and chills lasting for 5 days. He complained of severe asthenia. He came back two weeks ago from a travel in New York city, USA with a two days stay on the Long Island countryside. He did not note any insect bite during his travel. Jaundice was noticed on examination as well as slightly enlarged spleen and liver. Blood test showed an abnormal blood cell count with neutropenia (0·5 G/L), lymphopenia (0·3 G/L), thrombocytopenia (30×109/L) and anemia (haemoglobin of 91 g/L) with a low reticulocyte count (45 G/L). Blood test also showed increased C reactive protein (154 mg/L), raised ferritin (5953 ng/ml) and liver enzymes were elevated (ALT 56 UI/L (N<40), AST 68 UI/L (N<53), Alkaline phosphatase 213 UI/L (N<129), total bilirubin 48·9 μmol/L). Haptoglobin was undetectable, LDH levels increased 620 UI/L (N<225) and coagulation tests were in normal range. Routine blood cultures were negative. Bone marrow aspiration was performed because of the severe cytopenias associated with a low reticulocyte count and showed typical hemophagocytosis . Microscopic examination of the blood smear and the bone marrow stained with Giemsa showed intra-eryhtrocytic parasites . The parasitemia was evaluated about 3% of red blood cells. Rapid testing for Plasmodium falciparum by PCR revealed negative. Intravenous treatment for babesiosis with clindamycin and quinine was started. Patient's clinical status improved with the resolution of the fever after 3 days of treatment. Blood analysis improved after the 10 days treatment course (haemoglobin 104 g/L, platelets 220×109/L, neutrophils 3 G/L, C reactive protein 16 mg/L). PCR testing on blood for babesia was positive . Babesia microti was confirmed by PCR with specific primers (LDH gene amplification) as shown on Figure B. Serology for Borrelia burgdorferi, the agent of Lyme disease, was negative.

Generated summary

Patient Presentation: An 82-year-old man from the French Riviera, with recent travel history to New York City and Long Island, presented with fever, chills, and severe asthenia lasting 5 days.

Clinical Presentation: The patient exhibited jaundice, and examination revealed slight splenomegaly and hepatomegaly. Blood tests indicated neutropenia, lymphopenia, thrombocytopenia, anemia, elevated liver enzymes, increased C-reactive protein, raised ferritin, undetectable haptoglobin, and elevated LDH.

Diagnosis: Bone marrow aspiration showed hemophagocytosis, and blood smear revealed intra-erythrocytic parasites with 3% parasitemia. PCR testing for Plasmodium falciparum was negative, but Babesia microti was confirmed by PCR.

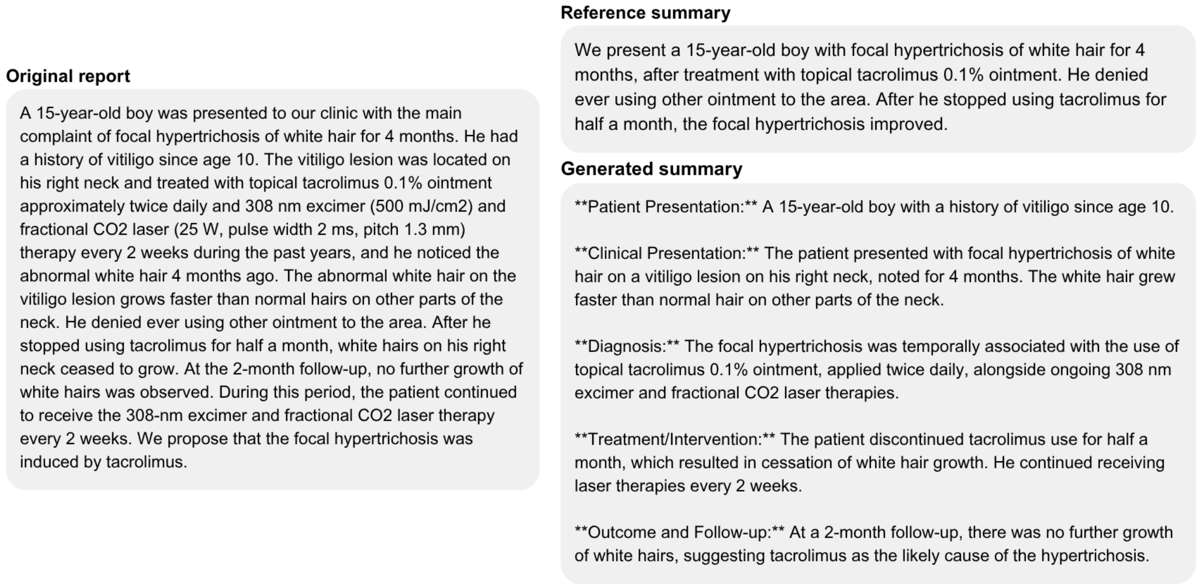
Treatment/Intervention: The patient received intravenous clindamycin and quinine for babesiosis.

Outcome and Follow-up: The patient's fever resolved after 3 days of treatment, and blood parameters improved significantly after a 10-day treatment course. PCR testing confirmed Babesia microti, and serology for Borrelia burgdorferi was negative.

Figure 5: Example of an original report next to the generated summary using the ISP technique, in which the generated summary covers key aspects in the original report.

Generally, the summaries tend to cover key features in the clinical report and contain logical paraphrasing, which likely led to the high BERTscores, i.e., a high semantic overlap. Figure 5 shows how the generated summary covers key aspects about the patient, such as the age (82), gender (male), and the travel history that is relevant for his symptoms. The summary then describes the symptoms (jaundice, etc.), and the tests and results that have led to a particular diagnosis, followed by the treatment and outcome. Based on a qualitative analysis with a small sample, we report that most generated summaries appear to be well-structured according to key features present in human-written clinical reports and consistent with the original report, which confirms the high BERTscores.

Comparing the generated summaries during the prompt updating phase and the reference summaries in the training set, we observe that the generated summaries tend to be longer than the gold-standard summaries. The generated summaries tend to be more detailed, including details about the specific tests done and the outcomes of these to reach a particular diagnosis. Figure 6 showcases an example in which the reference summary is much shorter than the LLM-generated summary. The generated summary contains more details regarding the different treatments that were done previously, while the reference summary focuses on the main complaint of the patient (“focal hypertrichosis of white hair") and the treatment that contributed to the patient’s improvement (“discontinued tacrolimus use"). This suggests that the LLM struggles with discerning key events from details that might be redundant for domain experts, who may be able to infer what procedures were done from a short and dense summary. In addition, the generated summaries always introduce the full form of abbreviations, which is not always the case in gold-standard summaries. These differences might have led to a lower ROUGE-L score.



Original report

A 15-year-old boy was presented to our clinic with the main complaint of focal hypertrichosis of white hair for 4 months. He had a history of vitiligo since age 10. The vitiligo lesion was located on his right neck and treated with topical tacrolimus 0.1% ointment approximately twice daily and 308 nm excimer (500 mJ/cm2) and fractional CO2 laser (25 W, pulse width 2 ms, pitch 1.3 mm) therapy every 2 weeks during the past years, and he noticed the abnormal white hair 4 months ago. The abnormal white hair on the vitiligo lesion grows faster than normal hairs on other parts of the neck. He denied ever using other ointment to the area. After he stopped using tacrolimus for half a month, white hairs on his right neck ceased to grow. At the 2-month follow-up, no further growth of white hairs was observed. During this period, the patient continued to receive the 308-nm excimer and fractional CO2 laser therapy every 2 weeks. We propose that the focal hypertrichosis was induced by tacrolimus.

Reference summary

We present a 15-year-old boy with focal hypertrichosis of white hair for 4 months, after treatment with topical tacrolimus 0.1% ointment. He denied ever using other ointment to the area. After he stopped using tacrolimus for half a month, the focal hypertrichosis improved.

Generated summary

Patient Presentation: A 15-year-old boy with a history of vitiligo since age 10.

Clinical Presentation: The patient presented with focal hypertrichosis of white hair on a vitiligo lesion on his right neck, noted for 4 months. The white hair grew faster than normal hair on other parts of the neck.

Diagnosis: The focal hypertrichosis was temporally associated with the use of topical tacrolimus 0.1% ointment, applied twice daily, alongside ongoing 308 nm excimer and fractional CO2 laser therapies.

Treatment/Intervention: The patient discontinued tacrolimus use for half a month, which resulted in cessation of white hair growth. He continued receiving laser therapies every 2 weeks.

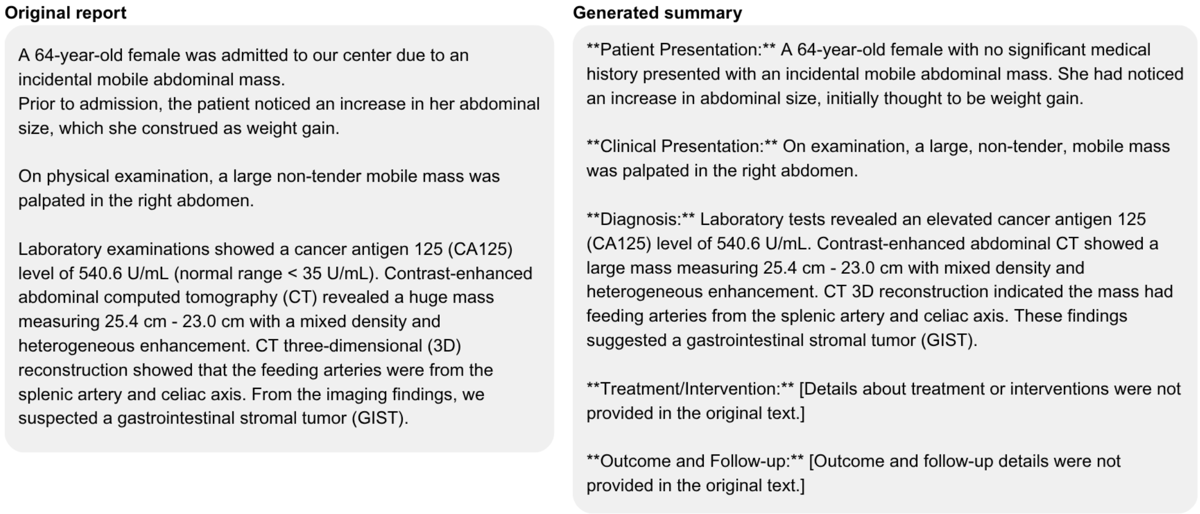
Outcome and Follow-up: At a 2-month follow-up, there was no further growth of white hairs, suggesting tacrolimus as the likely cause of the hypertrichosis.

Figure 6: Example of output during prompting procedure in which the generated summary is much longer and detailed than the reference summary.

Furthermore, we found that our prompt design resulted in the LLM including section headers with every generated summary (see Figure 6 and Figure 7). This led to the generated summaries strictly adhering to the structure as provided in the prompt, e.g., “Patient presentation”, “Diagnosis”, and “Treatment”. This explicit structure was generally absent in gold-standard summaries in the training set. Moreover, the strict adherence to the provided structure might have led to an ordering of key events and entities that is different from the reference summary. While the structure is logical and follows key features of clinical reports and summaries, it might have negatively impacted the structural overlap between the generated summary and the reference summary. Therefore, in this case, the inclusion of the section headers and strict adherence to the structure provided in the prompt might have contributed to the overall low ROUGE-L score.

Surprisingly, some generated summaries (12 texts in the test set, out of 3,396) are longer than the original texts (excluding the section headers). Examining the original clinical reports of this sample, it appears that the original reports are already quite brief and information-dense. The original texts of these generated summaries average around 135 words, which is much shorter than the average of 527 words in the entire test set. Close analysis of this sample reveals that the generated texts are mostly a repetition of the original text rather than a summary of key aspects and events. Figure 7 depicts how the generated summary copies phrases and sentences literally from the original report, only swapping a few verbs with close synonyms, e.g., “revealed” instead of “showed”.

The generated summaries also frequently include observations about missing information in the report, such as “The case report does not provide specific details on the outcome or follow-up. Typically,



Original report

A 64-year-old female was admitted to our center due to an incidental mobile abdominal mass.
Prior to admission, the patient noticed an increase in her abdominal size, which she construed as weight gain.

On physical examination, a large non-tender mobile mass was palpated in the right abdomen.

Laboratory examinations showed a cancer antigen 125 (CA125) level of 540.6 U/mL (normal range < 35 U/mL). Contrast-enhanced abdominal computed tomography (CT) revealed a huge mass measuring 25.4 cm - 23.0 cm with a mixed density and heterogeneous enhancement. CT three-dimensional (3D) reconstruction showed that the feeding arteries were from the splenic artery and celiac axis. From the imaging findings, we suspected a gastrointestinal stromal tumor (GIST).

Generated summary

Patient Presentation: A 64-year-old female with no significant medical history presented with an incidental mobile abdominal mass. She had noticed an increase in abdominal size, initially thought to be weight gain.

Clinical Presentation: On examination, a large, non-tender, mobile mass was palpated in the right abdomen.

Diagnosis: Laboratory tests revealed an elevated cancer antigen 125 (CA125) level of 540.6 U/mL. Contrast-enhanced abdominal CT showed a large mass measuring 25.4 cm - 23.0 cm with mixed density and heterogeneous enhancement. CT 3D reconstruction indicated the mass had feeding arteries from the splenic artery and celiac axis. These findings suggested a gastrointestinal stromal tumor (GIST).

Treatment/Intervention: [Details about treatment or interventions were not provided in the original text.]

Outcome and Follow-up: [Outcome and follow-up details were not provided in the original text.]

Figure 7: Example of an instance in the test set where the generated summary is longer than and very similar to the original report.

such a patient would require close monitoring and treatment adjustments based on laboratory and clinical responses." Figure 7 exemplifies this. This likely negatively impacted the automatic evaluation scores, but might be useful for domain experts who, in this way, can gain insight into what key information is missing in the original report.

5. Discussion

5.1. Prompt-Driven Model Behavior

With the iteratively updated prompt, the model exhibited the following behaviors:

- **Limited ability to compress content,** especially in the *Treatment* and *Patient Presentation* sections. This may be due to the lack of content filtering—the model tends to treat all information equally, failing to prioritise critical conditions and key treatments.
- **Structure-guided prompting may induce hallucinations or additional content,** particularly due to the decoder-only architecture. For example, if the prompt asks the model to summarize the *Outcome*, but the original text lacks such content, the model may fabricate information like "regular follow-up was scheduled." It may also lead to the model filling the gaps with statements such as "[Outcome and follow-up details are not provided in the original text."
- **Limited structural flexibility.** The model tends to follow the prompt-defined structure too rigidly, often generating key sentences by copying large portions of the original text with only minor adjustments. It is also prone to explicitly including section headings based on the focus points specified in the prompt, which may negatively affect ROUGE-L performance, particularly when the generated order differs from the gold reference.

Future work should focus on instructing the model to avoid redundant restatement, introducing counterfactual constraints to reduce hallucinated content, and developing more flexible structural control, such as preserving original abbreviations instead of expanding them, and softening the enforcement of fixed section headers.

5.2. Reflections on Evaluation Metrics

Although the original evaluation metric, ROUGE-L, effectively captures lexical overlap between the generated and gold summaries, it has notable limitations and may underestimate summary quality in certain cases. This is primarily because ROUGE-L is highly sensitive to variations in sentence structure and phrasing.

Furthermore, the manually annotated gold summaries are often highly compressed, frequently using abbreviations and omitting connective phrases. In contrast, the synthetic summaries tend to resemble patient-facing clinical reports, featuring more complete and explicit expressions. As a result, the two types of summaries may differ more in style than in substance. A low ROUGE-L score does not necessarily indicate poor summary quality, as the generated version may convey equivalent medical content in a different form.

Future work could incorporate metrics that account for structural coverage or introduce clinically grounded factual checks as complementary evaluation strategies.

6. Conclusions and Future Work

For this shared task on multilingual clinical document summarisation, we used perspective-aware iterative-self prompting (ISP) on LLMs via GPT4/4o, with the inspiration of the work by [4]. During the model development, we designed the following perspectives for summarisation, including Patient Presentation, Clinical Presentation, Diagnosis, Treatment, and Outcome (Follow-up). In conclusion, the perspective-aware iterative self-prompting (**PA-ISP**) on LLMs can help summarise lengthy clinical documents into short summaries while keeping the essence of the clinical knowledge, thus to help clinicians understand patients' healthcare history more efficiently, and help the patients to understand their condition better. The future work will also include lay/plain language adaptation into the summarisation so that patients with low health literacy level can better understand the clinical records, thus to improve the communications between patients and healthcare providers for better shared decision making. Local LLMs will be explored and trained for better privacy preservation. LLM Explainability and Reasoning are also our ongoing work. In addition, we plan to consider other languages such as the Spanish data from the shared task, as well as comparing more diverse prompts.

Acknowledgments

We thank Ida Korfage, Associate Professor at Erasmus MC and co-PI from the 4D Picture Project, for the valuable comments and revision on the Conference Abstract version of this paper. We thank the anonymous reviewers for their feedback on improving our article.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4o in order to: Grammar and spelling check.

References

[1] A. M. Stiggelbout, A. H. Pieterse, J. C. De Haes, Shared decision making: concepts, evidence, and practice, Patient education and counseling 98 (2015) 1172–1179.

[2] A. Stiggelbout, I. Griffioen, J. Brands, M. Melles, J. Rietjens, M. Kunneman, M. Van Der Kolk, C. Van Eijck, D. Snelders, Metro mapping: development of an innovative methodology to co-design care paths to support shared decision making in oncology, BMJ evidence-based medicine 28 (2023) 291–294.

[3] A. Nentidis, G. Katsimpras, A. Krithara, M. Krallinger, M. Rodríguez-Ortega, E. Rodriguez-López, N. Loukachevitch, A. Sakhovskiy, E. Tutubalina, D. Dimitriadis, G. Tsoumakas, G. Giannakoulas, A. Bekiaridou, A. Samaras, F. N. Maria Di Nunzio, Giorgio, S. Marchesin, M. Martinelli, G. Silvello, G. Paliouras, Overview of BioASQ 2025: The thirteenth BioASQ challenge on large-scale biomedical

semantic indexing and question answering, in: L. P. A. G. S. d. H. J. M. F. P. P. R. D. S. G. F. N. F. Jorge Carrillo-de Albornoz, Julio Gonzalo (Ed.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), 2025.

[4] P. Romero, L. Ren, L. Han, G. Nenadic, The manchester bees at peranssumm 2025: Iterative self-prompting with claude and o1 for perspective-aware healthcare answer summarisation, in: Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health), 2025, pp. 340–348.

[5] T. Naumann, A. Ben Abacha, S. Bethard, K. Roberts, D. Bitterman (Eds.), Proceedings of the 6th Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Mexico City, Mexico, 2024. URL: https://aclanthology.org/2024.clinicalnlp-1.0/.

[6] N. Collier, P. Ruch, A. Nazarenko (Eds.), Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP), COLING, Geneva, Switzerland, 2004. URL: https://aclanthology.org/W04-1200/.

[7] D. Demner-Fushman, S. Ananiadou, M. Miwa, K. Roberts, J. Tsujii (Eds.), Proceedings of the 23rd Workshop on Biomedical Natural Language Processing, Association for Computational Linguistics, Bangkok, Thailand, 2024. URL: https://aclanthology.org/2024.bionlp-1.0/.

[8] L. Han, S. Gladkoff, G. Erofeev, I. Sorokina, B. Galiano, G. Nenadic, Neural machine translation of clinical text: an empirical investigation into multilingual pre-trained language models and transfer-learning, Frontiers in Digital Health 6 (2024) 1211564.

[9] Z. Ling, Z. Li, P. Romero, L. Han, G. Nenadic, Malei at the plaba track of tac-2024: Roberta for task 1–llama3. 1 and gpt-4o for task 2, PLABA at TREC 2024 (2025).

[10] Z. Li, S. Belkadi, N. Micheletti, L. Han, M. Shardlow, G. Nenadic, Investigating large language models and control mechanisms to improve text readability of biomedical abstracts, in: 2024 IEEE 12th International Conference on Healthcare Informatics (ICHI), IEEE, 2024, pp. 265–274.

[11] S. Belkadi, L. Han, Y. Wu, G. Nenadic, Exploring the value of pre-trained language models for clinical named entity recognition, in: 2023 IEEE International Conference on Big Data (BigData), IEEE, 2023, pp. 3660–3669.

[12] H. Tu, L. Han, G. Nenadic, Extraction of medication and temporal relation from clinical text using neural language models, in: 2023 IEEE International Conference on Big Data (BigData), IEEE, 2023, pp. 2735–2744.

[13] Y. Cui, L. Han, G. Nenadic, Medtem2. 0: Prompt-based temporal classification of treatment events from discharge summaries, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), 2023, pp. 160–183.

[14] P. Romero, L. Han, G. Nenadic, Medication extraction and entity linking using stacked and voted ensembles on llms, in: Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health), 2025, pp. 303–315.

[15] L. Ren, S. Belkadi, L. Han, W. Del-Pinto, G. Nenadic, Synthetic4health: Generating annotated synthetic clinical letters, Frontiers in Digital Health 7 (2025) 1497130.

[16] L. Ren, S. Belkadi, L. Han, W. Del-Pinto, G. Nenadic, Beyond reconstruction: generating privacy-preserving clinical letters, in: Proceedings of the Sixth Workshop on Privacy in Natural Language Processing, 2025, pp. 60–74.

[17] A. Paul, D. Shaji, L. Han, W. Del-Pinto, G. Nenadic, Deidclinic: A multi-layered framework for de-identification of clinical free-text data, arXiv preprint arXiv:2410.01648 (2024).

[18] S. Agarwal, M. S. Akhtar, S. Yadav, Overview of the PerAnsSumm 2025 shared task on perspective-aware healthcare answer summarization, in: S. Ananiadou, D. Demner-Fushman, D. Gupta, P. Thompson (Eds.), Proceedings of the Second Workshop on Patient-Oriented Language Processing (CL4Health), Association for Computational Linguistics, Albuquerque, New Mexico, 2025, pp. 445–455. URL: https://aclanthology.org/2025.cl4health-1.41/. doi:`10.18653/v1/2025.cl4health-1.41`.

[19] M. Rodríguez-Ortega, E. Rodríguez-Lopez, S. Lima-López, C. Escolano, M. Melero, L. Pratesi, L. Vigil-Gimenez, L. Fernandez, E. Farré-Maduell, M. Krallinger, Overview of MultiClinSum task

at BioASQ 2025: evaluation of clinical case summarization strategies for multiple languages: data, evaluation, resources and results., in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), CLEF 2025 Working Notes, 2025.

A. MaLei Team Online Resources

The sources for the MaLei Team at MultiClinSUM shared task 2025 will be available via

- GitHub https://github.com/Libo-Ren/MultiClinSum,
- Our earlier PerAnSumm page from Manchester Bees https://github.com/pabloRom2004/-PerAnsSumm-2025.

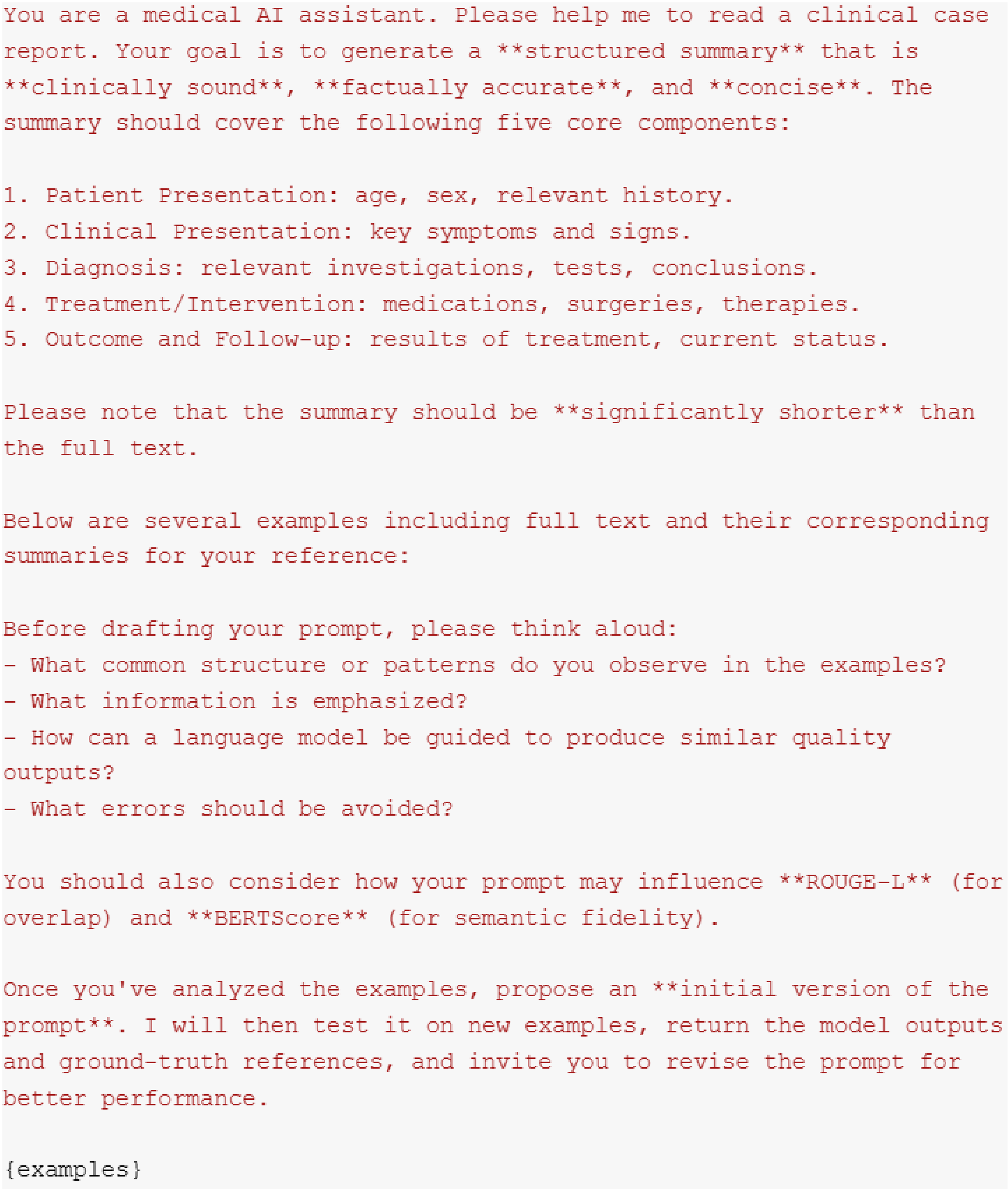
B. Prompt

B.1. Start Prompt

An example of starting a prompt using ISP is shown in Figure 8.

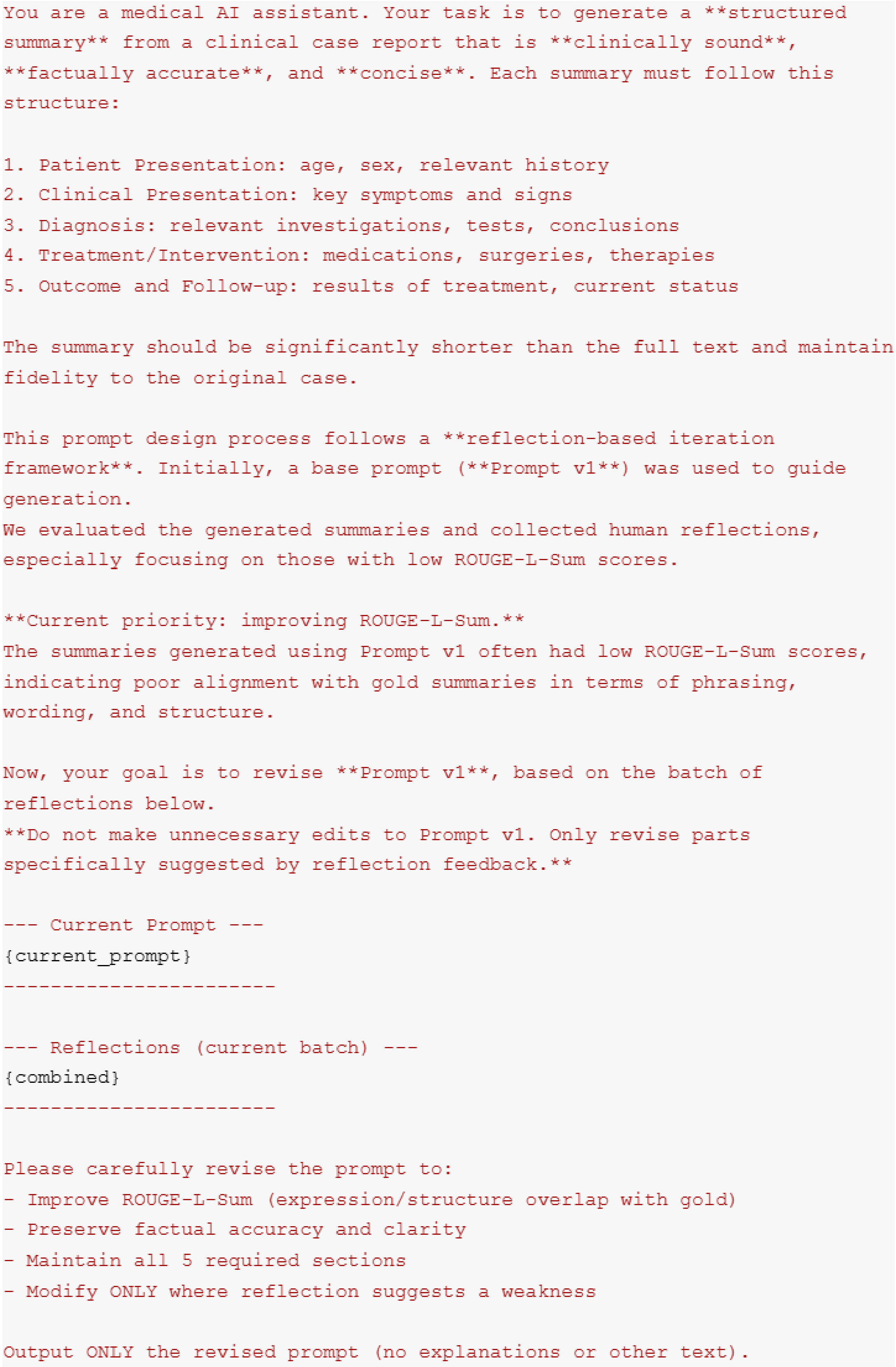
B.2. Instruction for Prompt Refinement

Meta-prompt used to instruct the LLM to revise the summary-generation prompt is shown in Figure 9.



```
You are a medical AI assistant. Please help me to read a clinical case
report. Your goal is to generate a **structured summary** that is
**clinically sound**, **factually accurate**, and **concise**. The
summary should cover the following five core components:

1. Patient Presentation: age, sex, relevant history.
2. Clinical Presentation: key symptoms and signs.
3. Diagnosis: relevant investigations, tests, conclusions.
4. Treatment/Intervention: medications, surgeries, therapies.
5. Outcome and Follow-up: results of treatment, current status.

Please note that the summary should be **significantly shorter** than
the full text.

Below are several examples including full text and their corresponding
summaries for your reference:

Before drafting your prompt, please think aloud:
- What common structure or patterns do you observe in the examples?
- What information is emphasized?
- How can a language model be guided to produce similar quality
outputs?
- What errors should be avoided?

You should also consider how your prompt may influence **ROUGE-L** (for
overlap) and **BERTScore** (for semantic fidelity).

Once you've analyzed the examples, propose an **initial version of the
prompt**. I will then test it on new examples, return the model outputs
and ground-truth references, and invite you to revise the prompt for
better performance.

{examples}
```

Figure 8: An example of starting prompt



```
You are a medical AI assistant. Your task is to generate a **structured
summary** from a clinical case report that is **clinically sound**,
**factually accurate**, and **concise**. Each summary must follow this
structure:

1. Patient Presentation: age, sex, relevant history
2. Clinical Presentation: key symptoms and signs
3. Diagnosis: relevant investigations, tests, conclusions
4. Treatment/Intervention: medications, surgeries, therapies
5. Outcome and Follow-up: results of treatment, current status

The summary should be significantly shorter than the full text and maintain
fidelity to the original case.

This prompt design process follows a **reflection-based iteration
framework**. Initially, a base prompt (**Prompt v1**) was used to guide
generation.
We evaluated the generated summaries and collected human reflections,
especially focusing on those with low ROUGE-L-Sum scores.

**Current priority: improving ROUGE-L-Sum.**
The summaries generated using Prompt v1 often had low ROUGE-L-Sum scores,
indicating poor alignment with gold summaries in terms of phrasing,
wording, and structure.

Now, your goal is to revise **Prompt v1**, based on the batch of
reflections below.
**Do not make unnecessary edits to Prompt v1. Only revise parts
specifically suggested by reflection feedback.**

--- Current Prompt ---
{current_prompt}
-----------------------

--- Reflections (current batch) ---
{combined}
-----------------------

Please carefully revise the prompt to:
- Improve ROUGE-L-Sum (expression/structure overlap with gold)
- Preserve factual accuracy and clarity
- Maintain all 5 required sections
- Modify ONLY where reflection suggests a weakness

Output ONLY the revised prompt (no explanations or other text).
```

Figure 9: Instruction for Prompt Refinement