Experiments in Automated Generation of Discharge Summaries in Italian

Lorenzo Ruinelli[1,2], Amos Colombo[1,2], Mathilde Rochat[3,4], Sotirios Georgios Popeskou[4,5], Andrea Franchini[6], Sandra Mitrović[6], Oscar Lithgow[6], Joseph Cornelius[6], Fabio Rinaldi[6]

[1] Team Innovation & Research, Area ICT,
Ente Ospedaliero Cantonale (EOC), Bellinzona, Switzerland
[2]Clinical Trial Unit, Ente Ospedaliero Cantonale (EOC), Bellinzona, Switzerland
[3] Servizio di Medicina Interna, Ospedale Regionale di Lugano,
Ente Ospedaliero Cantonale (EOC), Bellinzona, Switzerland
[4]Facoltà di Scienze Biomediche, Università della Svizzera Italiana (USI), Lugano, Switzerland
[5] Dipartimento di Chirurgia Viscerale, Ospedale Regionale di Lugano,
Ente Ospedaliero Cantonale (EOC), Bellinzona, Switzerland
[6] Dalle Molle Institute for AI (IDSIA - USI/SUPSI), Lugano, Switzerland
{lorenzo.ruinelli,amos.colombo,mathilde.rochat,sotiriosgeorgios.popeskou}@eoc.ch
{andrea.franchini,sandra.mitrovic,oscarwilliam.lithgow,joseph.cornelius,fabio.rinaldi}@idsia.ch

Abstract

Hospital discharge letters are a fundamental component of patient management, as they provide the crucial information needed for patient post-hospital care. However their creation is very demanding and resource intensive, as it requires consultation of several reports documenting the patient's journey throughout their hospital stay. Given the increasing pressures on doctor's time, tools that can draft a reasonable discharge summary, to be then reviewed and finalized by the experts, would be welcome. In this paper we present a comparative study exploring the possibility of automatic generation of discharge summaries within the context of an hospital in an Italian-speaking region and discuss quantitative and qualitative results. Despite some shortcomings, the obtained results show that a generic generative system such as ChatGPT is capable of producing discharge summaries which are relatively close to the human generated ones, even in Italian.

Keywords: Generating Discharge Summaries, LLMs, NLP

1. Introduction

The management of an hospitalization foresees the preparation of a Discharge Letter (DL) to summarize important information about the patient's diagnosis, treatment, medications, follow-up care, and any additional instructions or recommendations for the patient's ongoing health management. The primary goal of a DL is to convey critical information regarding a patient's care and treatment throughout their hospitalization to their general practitioner or primary care provider. The redaction of DLs is a resource-intensive process, both for the caretaker and the hospital (Golder et al., 2011; Cocco, 2012). The process often involves junior physicians who initially compose the first draft, which is then reviewed and validated by senior physicians before finalization. Physicians incur high risks of burnout (Hartman et al., 2023), which has been correlated to the bureaucratic tasks involved in their daily activities (Reith, 2018). While certain sections of the letter necessitate straightforward data extraction from the clinical records, others call for the capacity to distill and summarize complex clinical notes effectively. To fully or partially automate this process would imply a reduction in the time investment from the physician (Reith, 2018).

This paper explores the potential of large language models (LLMs) in enhancing the summarization of clinical records, written in Italian. In particular, we present an experiment aimed at validating the effectiveness of utilizing LLMs for supporting the summarizing of clinical diaries to be integrated into the discharge letter. The experiment is grounded in real-world clinical diaries correlated with their associated discharge letters, which are provided by our partner hospital. The evaluation process involves expert knowledge assessment and similarity-based metrics, with the aim of comparing the quality of the summaries generated by the LLM against the manually generated summaries (i.e. the DL).

2. Related Work

The interest and relevance of the task of automated generation of discharge summaries is shown by several publications and initiatives such as the BioNLP ACL'24 Shared Task on Streamlining Discharge Documentation (Xu, 2024). The generation of discharge summaries specifically tailored

Proceedings of CL4Health 2024 @LREC-COLING-2024, pages 137–144
20 May, 2024.

to the needs of the patient, aiming to maximize readability and understandability without sacrificing correctness, is discussed by (Zaretsky et al., 2024; Eppler et al., 2023). Other projects, as in (Ando et al., 2022; Hartman et al., 2023), research better strategies for summarizing structured or unstructured medical notes while still maintaining the domain's expert terminology, akin to our own goal. Given the recent advancements in transformer-based architecture and their performance in text summarization, recent studies almost exclusively rely on transformer-based neural network architecture for their experiments, such as (Ando et al., 2022) with BERT, (Hartman et al., 2023) with BERT and BART. Studies such as (Zaretsky et al., 2024; Eppler et al., 2023) approach the problem of text summarization through the use of readily available LLMs, specifically GPT-4.0. The problem then revolves around enhancing the language generation model by providing instructions to the LLM about the task, also known as prompt-engineering. The latter is a heuristic process highly specific to its target model. The use of more tailored prompts in these studies has shown measurable improvements in most metrics.

The typical evaluation strategies we find in the literature often involve the following metrics: ROUGE (Lin and Hovy, 2003), BLEU (Papineni et al., 2002), BertScore (Zhang et al., 2019), BLEURT (Sellam et al., 2020) and MoverScore (Zhao et al., 2019), which score the similarity between documents, usually between the reference, written by a physician, and the generated one. Some studies employ ROUGE and BLEU (Ando et al., 2022; Hartman et al., 2023), now considered less sophisticated than their neural network alternatives, which offer a more human-like judgment. Neural network-based metrics usually consider semantic and contextual information, thus providing more reliable insight into the generated text when comparing it against the reference, as employed by (Ando et al., 2022). Some authors, as (Hartman et al., 2023; Zaretsky et al., 2024; Eppler et al., 2023) supplement their evaluations by involving one or more domain's experts to review the generated document and provide a correctness measure based on human judgment.

Related literature involving the usage of LLMs in the medical context with Italian language seems to be quite restricted, studying for example the capacity of LLMs (including ChatGPT-3.5 and ChatGPT-4) to answer the questions and provide templates related to structured reports in radiology (Mallio et al., 2023). Another study investigated ChatGPT potential in generating and annotating goal-oriented dialogues, and used as one of the use cases a scenario when doctor needs to explain the diagnosis and treatment to a patient (Labruna et al., 2023). In (Montagna et al., 2023), a comprehensive framework for creating an LLM-based chatbot system that assists chronic patients is introduced. To the best of our knowledge, this is the first study focusing on discharge letters/summaries in Italian.

3. Methodology

We screened the hospital database and collected both discharge letters and the corresponding clinical notes utilized in their composition. Clinical notes are written by nurses and doctors during the patient's stay, describing the current status of the patient and the future steps in the patients care. The timeframe was restricted to a recent six-month period. The language of discharge letters and corresponding clinical notes is Italian. Our focus was on simple cases, defined as clinical notes with a character length ranging between 3400 and 4000. This character length was chosen in order to not exceed the ChatGPT character limit and is close to the mean length of the clinical notes. Additionally, we targeted two medical specialties: surgery and medicine, sampling 30 cases from each group. Discharge letters from medicine cases tend to be more complex in nature compared to those of surgical cases. Clinical notes and discharge letters were de-identified using an internally developed tool capable of removing patient names and ages, contacts, locations and organizations. We produced two summaries for each case: one utilizing ChatGPT-3.5 (denoted as $AI_{3.5}$) and the other using ChatGPT-4 (denoted as AI_4). As the purpose of this experiment was only to test the feasibility of the idea, we used a prompt composed by a simple request (*"Crea un riassunto del seguente decorso clinico da includere nella lettera di uscita"*)[1] followed by the clinical notes in JSON format[2]. Each case had then 4 documents: the prompt including the clinical notes (P), the summary written by medical doctors (MD), and the two summaries generated by ChatGPT ($AI_{3.5}$ and AI_4).

The similarities between each document pairs were evaluated the using following metrics:

- ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation) (Lin and Hovy, 2003), a recall-oriented metric based on longest shared common subsequence in the documents.
- BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002), a precision-based metric quantifying the overlap of n-grams between the documents

[1] In English: "Create a summary of the following clinical discourse to include in the discharge letter"

[2] An example is provided in the Appendix A.

- BERTscore (Zhang et al., 2019), which harnesses contextual embeddings from BERT to compute the similarity between the documents
- BLUERT (Sellam et al., 2020), a BERT-based text similarity evaluation metric modeled to mimic human judgment and optimized for generality. It is designed to compare sentences, so it might not be effective on entire documents.

We use ROUGE-L as the chosen ROUGE metric computed with the rouge_score python library. The BLEURT score was computed using the code from the official BLEURT GitHub repository. The standard scorer uses the BERT-Tiny[3] model.

Additionally, we conducted an expert evaluation of the ChatGPT-generated summaries with the assistance of two medical doctors.

4. Results

	Medicine		*Surgery*	
Metric	$AI_{3.5}$	AI_4	$AI_{3.5}$	AI_4
BERTscore	0.8890	0.9050	0.8900	0.8950
BLEU	0.0003	0.0002	0.0006	0.0006
BLEURT	-0.3710	-0.3740	-0.3420	-0.3240
ROUGE	0.1130	0.1280	0.1370	0.1360

Table 1: Comparison of DLs generated by doctors (MD) versus AI generated ones ($AI_{3.5}$ and AI_4) in the Medicine & Surgery specialties (30 cases each)

Table 1 illustrates the average of the numerical comparison for the 30 samples in each of the two specialties. It shows that ChatGPT-4 generally outperforms ChatGPT-3.5 across all metrics except for BLEU. Specifically, ChatGPT-4 has higher BERTscore and ROUGE scores for general medicine, indicating better semantic similarity and n-gram overlap with reference texts. Additionally, ChatGPT-4 achieves better BERTscore for surgery than ChatGPT-3.5. BLEU scores are very low for both models, but slightly better for ChatGPT-3.5 in one instance and equal in another. BLEURT scores, while negative for both, are slightly higher for ChatGPT-4, suggesting a slight improvement in semantic quality. Overall, ChatGPT-4 demonstrates a marginal but consistent improvement in text generation quality over ChatGPT-3.5.

The qualitative evaluation was conducted with the collaboration of two medical doctors, one specialized in Medicine and the other in Surgery. Each doctor evaluated cases from their respective specialty. Both doctors found the AI-generated summaries well-done and potentially useful, expressing a preference for those generated by ChatGPT-4 over ChatGPT-3.5. The following section aims to illustrate the problems that have been identified, using one case for each of the two specialties. Figure 1 refers to a general medicine case, while Figure 2 refers to a surgery case. Both figures show on top the original human-generated summary, and on the bottom the summary generated using GPT-4.

In the medicine case (Figure 1), we observe that a significant portion of the medical doctor's summary (highlighted in yellow) reports information that was not present in the clinical notes processed by the AI models. This is because this information comes from the notes collected in the emergency room, which were not used in our experiment. The AI-generated summary begins by stating when the patient was discharged (see 1 in Figure 1)[4]. While this information is factually correct, it does not follow the typical style of discharge letters, which typically do not begin in this manner. The AI-generated summaries also included a series of stay-related information that are not relevant in this discharge letter, namely: fever episodes (2), infusion treatment (3), addressing hypokalemia (6), conducting regular laboratory tests (8), planning the return home after the hospital stay (9), and treating with Sintrom due to INR values (10). The AI-generated summary includes a sentence stating that the patient was treated with azithromycin because of a positive result on the Legionella test (4). However, this is not entirely accurate, as the treatment decision was based on the positive test result, along with the patient's medical history and other diagnostic investigations. The AI models incorrectly interpreted the Italian acronym for vital parameters (PV) as venous pressure (5).

In the surgery case (Figure 2), we notice that the yellow-highlighted portion is smaller compared to the medicine case. This indicates that the notes used in our experiment include a larger portion of the necessary information. The AI models omitted two important pieces of information: that the patient had an intraductal papillary mucinous neoplasm (IPMN), as well as the result of the cholangiography. Similarly to the medicine case, the AI models included a series of stay-related information that are not relevant in a discharge letter: comprehensive blood tests (see 2 in Figure 2), a Cholangiography performed and report pending (3, 5), fasting blood glucose test (4), and a summary of the patient's status (7). The AI-generated summary also included a sentence (6) that is not entirely correct from a clinical perspective, and it also hallucinated about colestasi (1).

[3] https://github.com/google-research/bert

[4] Notice that specific text segments in the figures are identified by a superscript, which we use from here on to refer to them.

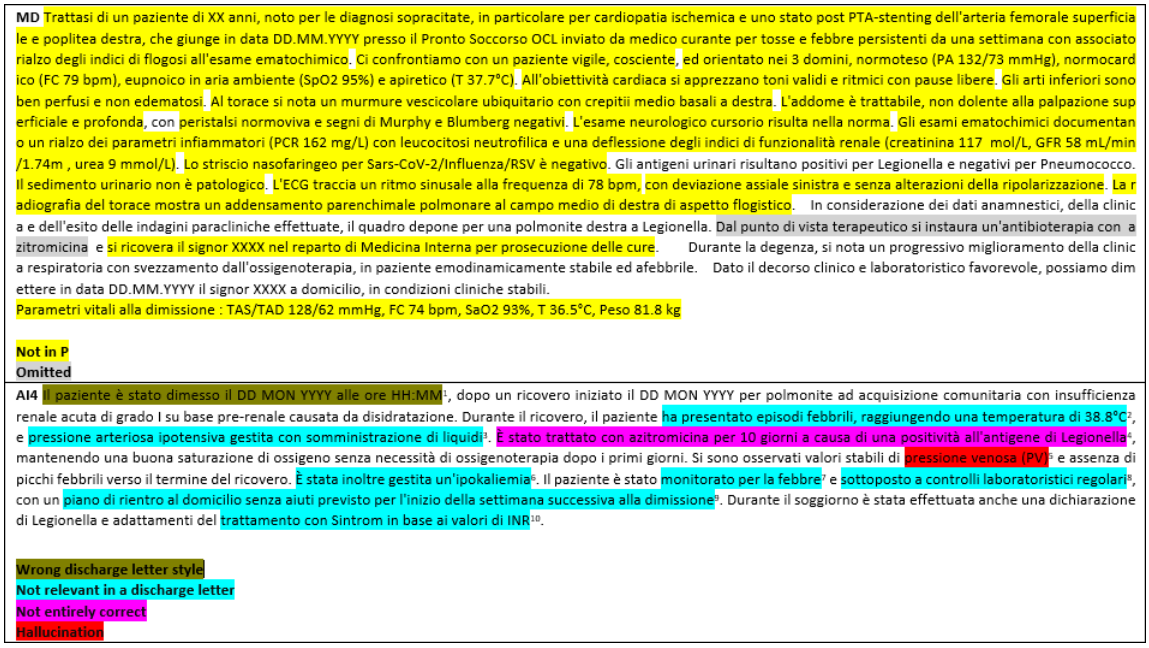

Figure 1: Comparison of the MD summary with the AI-generated summary for a medicine case. Color coding: yellow - information not present in clinical notes and consequently neither in the prompt; grey - information provided in the prompt but omitted in the generated summary; olive - wrong discharge letter style; cyan - information not relevant in a discharge letter; magenta - information not entirely correct; red - hallucination.

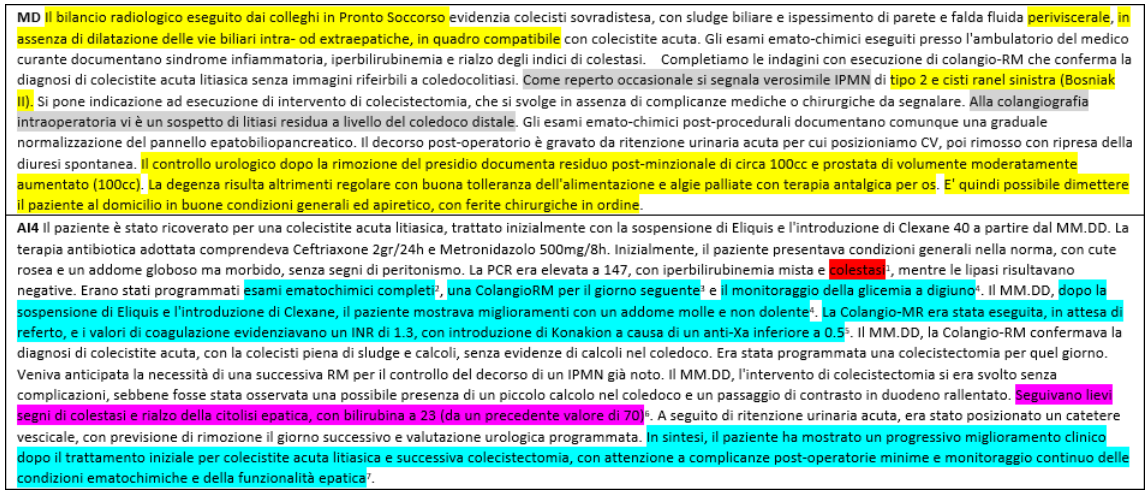

Figure 2: Comparison of the MD summary with the AI-generated summary for a surgery case. Color coding: yellow - information not present in the prompt; grey - information provided in the prompt but omitted in the generated summary; cyan - information not relevant in a discharge letter; magenta - information not entirely correct; red - hallucination.

5. Discussion

Our experiment shows that LLMs offer considerable potential for improving the summarization of clinical records in healthcare, particularly for the preparation of discharge letters. However, their adoption faces challenges.

First and foremost, our experiment setup, which relies on a publicly available ChatGPT model, raises significant privacy concerns, as the de-identification tool may fail to completely remove sensitive parts of the data. In our case, since the number of cases was limited, the risk was mitigated by manual validation of complete de-identification before submission. Secondly, it remains challenging to ensure that the generated summaries consistently maintain a high level of accuracy. The most common quantitative metrics excel at capturing the

broad meaning of a text, but they cannot capture specific details crucial in clinical practice. Qualitative evaluations, while providing more specific indications, rely on expert evaluation, which is often subjective, and also extremely expensive to obtain. The observation that the versions of the DLs generated by ChatGPT-4 were considered better than those generated by ChatGPT-3.5, in particular with enhanced understanding of temporal aspects, is a positive signal that indicates further improvements can be expected. From a quantitative standpoint, it's interesting to observe that ChatGPT-4 produces longer summaries compared to ChatGPT-3.5, with an increase of around 20%.

We would like to add several observations regarding the obtained quantitative results. First, as was evidenced in yellow coded parts in Figure 1 and Figure 2, given that AI models operated exclusively based on the information provided in the prompts, which were missing some of the extra information available to doctors, the content of expert summary (MD) remarkably extends that of the AI-generated one. This clearly drastically reduces the overlapping parts of AI-generated DL and MD summaries. Given that ROUGE-L is based on the longest common sequence of words (not necessarily consecutive, but still in order) shared between AI-generated DL and ground truth (MD summaries), it is thus not surprising that the obtained scores are very low for both ChatGPT-3.5 and ChatGPT-4. The same problem reflects even more drastically on BLEU scores, since they exploit *consecutive* sequences of words (in our case, up to 3-grams were considered). Finally, there are at least three reasons for obtaining somewhat unexpected negative BLEURT scores: 1) using BERT-Tiny as checkpoint was probably not the best option since although very light is also known to be very inaccurate[5]; 2) the more stable BLEURT checkpoint BLEURT-20 was not tested on Italian language; 3) BLEURT scores heavily depend on the quality and representatives of the training data and may not fully capture the nuances of language quality across different domains or contexts. We thus recommend to consider BLEURT scores with caution.

Clinical notes are very detailed in nature, as they must contain all the information utilized for patient management during hospitalization. In our experimental setup, the AI models appeared unable to accurately filter relevant information to be included in the discharge letter. To address this gap, we could modify the structure of the clinical notes (e.g. by implementing a more structured reporting format for the information), or enhance the prompt, or try different models.

Given the positive outcome of the feasibility study described in this paper (as corroborated also by medical experts), we are now setting up a larger and more advanced experiment which will enable us to tackle some of the shortcomings previously described. The first crucial step will be to use a local installation of an advanced open-source domain-specific model such as (Chen et al.; Jin et al.; Li et al., 2023), which were specifically trained on medical terminology and context. These specialized models can better capture the intricacies of medical causality, enhance the coherence and reduce errors in term interpretation. Additionally, the local installation will enable larger experiments, while at the same time mitigating privacy risks. The experimental strategy will involve a combination of prompt engineering techniques, including knowledge-infused prompting, chained inference, and corrective retrieval-augmented generation (Yan et al.). During *prompt engineering* (Brown et al., 2020), we can enrich the model's prompt with specific information about guidelines governing the generation process. In a *chained inference process* the AI model self-reflects and critiques its initial answer, subsequently generating a refined response based on this introspection. Finally, by contextualizing prompts with clinical topics from reputable sources, potentially obtained through *retrieval augmented generation*, we aim to provide more relevant and grounded knowledge to the LLM, enabling it to accurately correlate medical information.

6. Conclusion

In this paper we presented the results of a preliminary experiment aimed at testing the feasibility of automatic generation of discharge summaries in Italian. The setting of our experiment is deliberately oversimplified, in order to enable the validation of the idea, before attempting experiments that would require larger investments, such as the in-house installation and usage of an open-source LLM.

The results show that a generic generative system such as ChatGPT is capable of producing discharge summaries which are relatively close to the human generated ones, even in Italian. We have however noticed some shortcomings, which will need to be addressed in order for the system to be used in production. These observations have been collected and will guide the development of strategies to overcome them, such as enhanced prompting and retrieval-augmented generation.

[5] https://github.com/google-research/bleurt

7. Limitations and Ethical Considerations

We are aware that this work has several limitations. First, we operate with limited number of clinical notes and consider only two medical specialities. Second, we consider only Italian language hence the obtained insights might not be transferable to other languages.

In accordance with ethical principles, this scientific study exploits data de-identification to safeguard the privacy and confidentiality of patients, thus aiming to minimize the risk of potential harm or identification. All clinical notes were manually revised after de-identification, to make sure that no instance of personally identifiable information was left in them.

We also contacted the ethical committee and they confirmed that this type of research did not require their authorization.

8. Bibliographical References

Kenichiro Ando, Mamoru Komachi, Takashi Okumura, Hiromasa Horiguchi, and Yuji Matsumoto. 2022. Is in-hospital meta-information useful for abstractive discharge summary generation? In 2022 International Conference on Technologies and Applications of Artificial Intelligence (TAAI), pages 143–148. IEEE.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. Advances in neural information processing systems, 33:1877–1901.

Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. MEDITRON-70B: Scaling Medical Pretraining for Large Language Models.

Giuseppe Cocco. 2012. Bureaucracy and medicine, an unholy marriage. Cardiovasc Med, 15:243–244.

Michael B Eppler, Conner Ganjavi, J Everett Knudsen, Ryan J Davis, Oluwatobiloba Ayo-Ajibola, Aditya Desai, Lorenzo Storino Ramacciotti, Andrew Chen, Andre De Castro Abreu, Mihir M Desai, et al. 2023. Bridging the gap between urological research and patient understanding: the role of large language models in automated generation of layperson's summaries. Urology practice, 10(5):436–443.

Nikolaos Giarelis, Charalampos Mastrokostas, and Nikos Karacapilidis. 2023. Abstractive vs. extractive summarization: An experimental review. Applied Sciences, 13(13).

Lukas Golder, Claude Longchamp, Martina Imfeld, Silvia Ratelband-Pally, Stephan Tschöpe, Andreas Stettler, and Jonas Ph. Kocher. 2011. Drg: Befürchtungen einer zunehmenden bürokratisierung der medizin. Technical report, Gfs.bern, Hirschengraben 5,Postfach 6323, 3001 Bern, Switzerland.

Vince C Hartman, Sanika S Bapat, Mark G Weiner, Babak B Navi, Evan T Sholle, and Thomas R Campion Jr. 2023. A method to automate the discharge summary hospital course for neurology patients. Journal of the American Medical Informatics Association, 30(12):1995–2003.

Mingyu Jin, Qinkai Yu, Chong Zhang, Dong Shu, Suiyuan Zhu, Mengnan Du, Yongfeng Zhang, and Yanda Meng. 2024. Health-llm: Personalized retrieval-augmented disease prediction model. arXiv preprint arXiv:2402.00746.

Qiao Jin, Won Kim, Qingyu Chen, Donald C Comeau, Lana Yeganova, W John Wilbur, and Zhiyong Lu. MedCPT: Contrastive Pre-trained Transformers with large-scale PubMed search logs for zero-shot biomedical information retrieval. 39(11):btad651.

Tiziano Labruna, Sofia Brenna, Andrea Zaninello, and Bernardo Magnini. 2023. Unraveling chatgpt: A critical analysis of ai-generated goal-oriented dialogues and annotations. In International Conference of the Italian Association for Artificial Intelligence, pages 151–171. Springer.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023. LLaVA-Med: Training a Large Language-and-Vision Assistant for Biomedicine in One Day.

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics, pages 150–157.

Carlo A Mallio, Andrea C Sertorio, Caterina Bernetti, and Bruno Beomonte Zobel. 2023. Large language models for structured reporting in

radiology: performance of gpt-4, chatgpt-3.5, perplexity and bing. La radiologia medica, 128(7):808–812.

Sara Montagna, Stefano Ferretti, Lorenz Cuno Klopfenstein, Antonio Florio, and Martino Francesco Pengo. 2023. Data decentralisation of llm-based chatbot systems in chronic disease self-management. In Proceedings of the 2023 ACM Conference on Information Technology for Social Good, pages 205–212.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pages 311–318.

Thomas P Reith. 2018. Burnout in united states healthcare professionals: a narrative review. Cureus, 10(12).

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation. In ACL.

J. Xu. 2024. Discharge me: Bionlp acl'24 shared task on streamlining discharge documentation (version 1.2).

Ran Xu, Hejie Cui, Yue Yu, Xuan Kan, Wenqi Shi, Yuchen Zhuang, Wei Jin, Joyce Ho, and Carl Yang. 2023. Knowledge-Infused Prompting: Assessing and Advancing Clinical Text Data Generation with Large Language Models.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. Corrective Retrieval Augmented Generation.

Jonah Zaretsky, Jeong Min Kim, Samuel Baskharoun, Yunan Zhao, Jonathan Austrian, Yindalon Aphinyanaphongs, Ravi Gupta, Saul B Blecker, and Jonah Feldman. 2024. Generative artificial intelligence to transform inpatient discharge summaries to patient-friendly language and format. JAMA Network Open, 7(3):e240357–e240357.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance. arXiv preprint arXiv:1909.02622.

A. An Example of a Prompt

P Crea un riassunto del seguente decorso clinico da includere nella lettera di uscita:
[{"CONTENUTO": "Dimesso alle 14:00", "DATA": "2023-01-09 11:21:49"},
{"CONTENUTO": "Paziente stabile, PV nella norma. Apiretico."
"Procedere:"
"- sorveglianza clinica e febbre. Se persiste, valutare switch a levofloxacina"
"- labor controllo domani (K, INR)"
"- rientro al domicilio senza aiuti inizio settimana prossima",
"DATA": "2023-01-08 12:29:57"},
{"CONTENUTO": "Paziente stabile, PV nella norma. "
"Assenza di picchi febbrili. Buone Spo2 senza ossigenoterapia."
"Procedere:"
"- sorveglianza clinica e febbre. Se persiste, valutare switch a levofloxacina"
"- labor controllo lunedì (K, INR)"
"- rientro al domicilio senza aiuti inizio settimana prossima",
"DATA": "2023-01-07 10:23:32"},
{"CONTENUTO": "Paziente stabile, PV nella norma. Assenza di picchi febbrili. "
"Buone Spo2 senza ossigenoterapia. Per ipokaliemia in corso sostituzione."
"Procedere WEEKEND:"
"- sorveglianza clinica e febbre. Se persiste, valutare switch a levofloxacina"
"- labor controllo lunedì e RAD se stabilità clinica",
"DATA": "2023-01-06 10:01:10"},
{"CONTENUTO": "Febbrile durante la notte a 38°C, afebbrile durante il giorno. "
"Resto parametri nella notte. L: PCR stagnante, ipokaliemia. Non indicazione a switch antibioterapia"
"Procedere WEEKEND:"
"- sorveglianza clinica e febbre. Se persiste, valutare switch a levofloxacina"
"Procedere:"
"- RAD lunedì"
"- labor controllo lunedì",
"DATA": "2023-01-05 12:11:11"},
{"CONTENUTO": "Addendum 04.01.2022 "
"- Dischiarazione Legionella al xxxxxxxxxxxxxxx fatta oggi"
"- Sintrom da adattare secondo INR domani",
"DATA": "2023-01-04 18:33:03"},
{"CONTENUTO": "PV nella norma, afebbrile. Questa mattina saturava a 96% con 2L di O2, lo svezziamo e in giornata
“satura a 92% senza O2. "
"Cercare di mantenere senza O2."
"Procedere:"
"- Domani labor"
"- Aritromicina per 10 giorni"
"- RAD lunedì",
"DATA": "2023-01-04 14:57:55"},
{"CONTENUTO": "PV stabili, febbrile a 38.8°. No tosse. Crepitii mediobasali a destra. AG legionella positivi --> Th: Azitromicina
“10 giorni secondo evoluzione clinica “
“(per il momento impostata fino all"11, rivalutare). Schema sintron reimpostato"
"Procedere:"
"- Sorveglianza clinica e laboratoristica"
"- Se tutto Ok, RAD giovedì",
"DATA": "2023-01-03 12:37:34"},
{"CONTENUTO": "Paziente ricoverato per polmonite ad acquisizione comunitaria "
"con IRA AKIN I di origine pre renale su disidratazione. Vengo chiamata in serata per valori pressori ipotensivi "
"(PA 85/55 mmHg) in un paziente asintomatico. "
"Si somministrano 250 ml di liquidi in 30 minuti con risposta sulla pressione. Alla visita paziente vigile ed orientato. "
"Diaforetico e febbrile. Mucose secche.”
"Si velocizza la somministrazione dei liquidi prescrivendo 1000 ml in 12 ore.",
"DATA": "2023-01-03 00:02:17"}]

Figure 3: Example of a prompt (P)