Proceedings of Machine Learning Research 259:364–383, 2024 Machine Learning for Health (ML4H) 2024

Query-Guided Self-Supervised Summarization of Nursing Notes

Ya Gao [1] ya.gao@aalto.fi
Hans Moen [1] hans.moen@aalto.fi
Saila Koivusalo [2] saila.koivusalo@hus.fi
Miika Koskinen [2] miika.koskinen@hus.fi
Pekka Marttinen [1] pekka.marttinen@aalto.fi
[1] *Aalto University, Finland*
[2] *Helsinki University Hospital, Finland*

Abstract

Nursing notes, an important part of Electronic Health Records (EHRs), track a patient's health during a care episode. Summarizing key information in nursing notes can help clinicians quickly understand patients' conditions. However, existing summarization methods in the clinical setting, especially abstractive methods, have overlooked nursing notes and require reference summaries for training. We introduce QGSumm, a novel query-guided self-supervised domain adaptation approach for abstractive nursing note summarization. The method uses patient-related clinical queries for guidance, and hence does not need reference summaries for training. Through automatic experiments and manual evaluation by an expert clinician, we study our approach and other state-of-the-art Large Language Models (LLMs) for nursing note summarization. Our experiments show: 1) GPT-4 is competitive in maintaining information in the original nursing notes, 2) QGSumm can generate high-quality summaries with a good balance between recall of the original content and hallucination rate lower than other top methods. Ultimately, our work offers a new perspective on conditional text summarization, tailored to clinical applications.

Keywords: abstractive text summarization, nursing notes, self-supervised learning

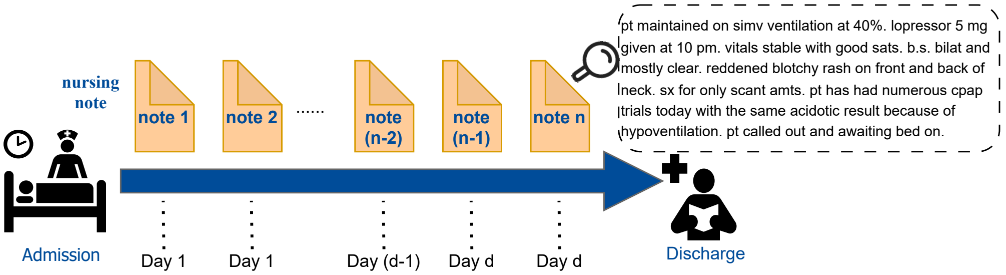
Data and Code Availability We use the publicly available database MIMIC-III (Johnson et al., 2016) in this work. Code is available in the repository.

Institutional Review Board (IRB) Since we used a public dataset in this work, our research does not require IRB approval.

1. Introduction

Nursing notes are important for clinicians to track a patient's health status and administered treatments during hospitalization (Törnvall and Wilhelmsson, 2008). However, a care episode may result in a large number of nursing notes, especially for patients suffering from complex health problems (Hall and Walton, 2004). Furthermore, the condensed nature of nursing notes makes them time-consuming for clinicians to understand (Clarke et al., 2013). Text summarization methods from Natural Language Processing (NLP) help distill the content of clinical notes (Wang et al., 2021), making crucial information more accessible and less time-consuming to review. This can be particularly beneficial for clinicians, and for conducting retrospective studies. These summarization techniques are divided into extractive (Pivovarov and Elhadad, 2015; Moen et al., 2016; Tang et al., 2019) and abstractive (Zhang et al., 2020b; Liu et al., 2022a; Searle et al., 2023). Extractive methods retain the original sentences and/or keyphrases but may lack coherence or fluency. On the other hand, abstractive methods produce smoother summaries but typically require explicit supervision, i.e., a reference summary as the ground truth, which are time-consuming to produce(O'Donnell et al., 2009).

Self-supervised abstractive text summarization (Chu and Liu, 2019; Elsahar et al., 2021) bypasses the problem of lacking ground-truth summaries. Previous self-supervised methods (e.g., Liu et al. (2021)) reduce the semantic distance between a summary and the original text. However, unconditionally making the semantic representations similar lacks control of the generated summaries and may result in missing relevant information, which can be important especially in clinical applications. Furthermore, the con-



pt maintained on simv ventilation at 40%. lopressor 5 mg given at 10 pm. vitals stable with good sats. b.s. bilat and mostly clear. reddened blotchy rash on front and back of neck. sx for only scant amts. pt has had numerous cpap trials today with the same acidotic result because of hypoventilation. pt called out and awaiting bed on.

Figure 1: From a patient's admission to discharge, multiple nursing notes may be generated. As shown in one artificial nursing note example, the notes could be poorly organized and lack clarity.

tent of nursing notes may be unclear, poorly organized and contain non-standard abbreviations (Figure 1). Hence, in the absence of supervised learning signals, guiding the model to generate a good summary becomes challenging.

In this paper, we propose a query-guided self-supervised domain adaptation framework for nursing note summarization, named QGSumm. A good summary of a clinical note is centered on the patient's condition. We achieve this by designing queries that concern the patient's condition, and we train the model to produce summaries such that answers to the queries based on the summaries are similar to the answers based on the original notes. This guidance makes our method highly suitable for the clinical field since the resulting summaries are centered on the information nurses and clinicians are most concerned about. To the best of our knowledge, our study is the first on abstractive summarization of nursing notes and on employing self-supervised learning for clinical note summarization.

Our primary contributions are: **(1)** The study focuses on nursing notes that play a critical role in clinical settings, filling a gap in previous research by introducing a method for abstractive nursing note summarization. Our method's ability to work without reference summaries highlights its practical applicability, while also being relatively light-weight and suitable for deployment in a secure environment with limited computing resources. **(2)** We propose a novel self-supervised domain adaptation framework. By leveraging patient-related queries, we guide the model to generate nursing note summaries that prioritize specific content, i.e., the patients' conditions and health status, without the need for manually written reference summaries as ground truth. **(3)** We conduct a comprehensive automated empirical study and a manual evaluation by an expert clinician, including state-of-the-art Large Language Models (LLMs), which have not been previously investigated for nursing note summarization. The experiments demonstrate especially our method's and GPT-4's ability to perform well in this task.

2. Related Work

Clinical Notes Summarization. Extractive summarization methods can preserve faithfulness but may have difficulties in maintaining coherence, and include earlier semantic similarity-based techniques (Pivovarov and Elhadad, 2015; Moen et al., 2016), as well as more recent attention-based methods (Tang et al., 2019; Reunamo et al., 2022; Kanwal and Rizzo, 2022). On the other hand, abstractive clinical notes summarization has been applied on discharge summaries (Shing et al., 2021; Adams et al., 2022; Searle et al., 2023), radiology reports (Zhang et al., 2020b; Van Veen et al., 2023; Nishio et al., 2024), doctor-patient conversations (Zhang et al., 2021; Krishna et al., 2021; Abacha et al., 2023) and problem lists (Gao et al., 2022, 2023). However, unlike ours, these works depend on data annotation or reference summaries. LLMs demonstrate a remarkable capability in clinical text understanding. Van Veen et al. (2024) extensively analyze the clinical text summarization performance of various LLMs with in-context learning (Lampinen et al., 2022) and QLoRA (Dettmers et al., 2024) adaptation.

Unsupervised and Self-Supervised Abstractive Text Summarization. The scarcity of annotated text has spurred interest in unsupervised and self-supervised text summarization. Previous works like source document reconstruction assume that a good

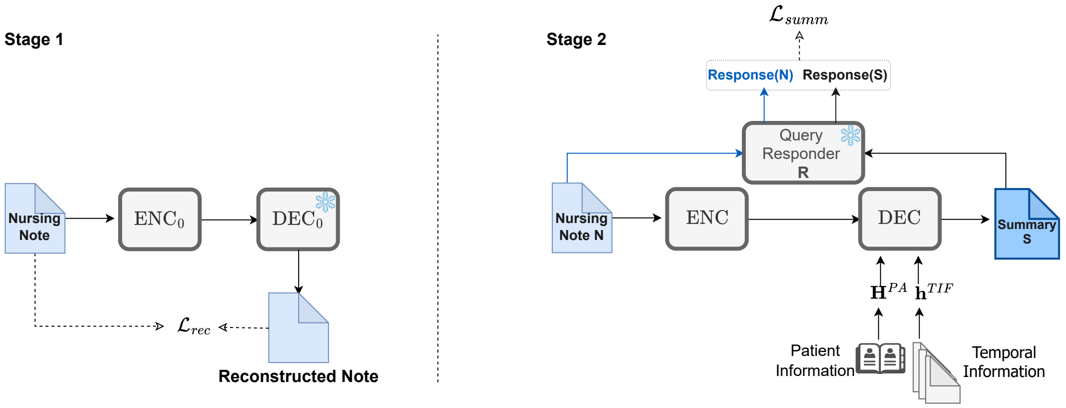


Figure 2: **Stage 1:** Fine-tune the encoder (ENC_0) by note reconstruction, where DEC_0 is frozen; **Stage 2:** Fine-tune the encoder (ENC) and decoder (DEC) in the self-supervised manner, where the query responder R is frozen. Raw patient information is processed by ENC into the embeddings $\mathbf{H}^{PA}$, as shown in Fig. 3

summary can reconstruct the source document (Chu and Liu, 2019), but such summaries may lack specific information due to limited control. Two-step approaches extract important information first and then perform abstractive summarization based on this extraction (Zhong et al., 2022; Ke et al., 2022; Liu et al., 2022b). However, the quality of such summaries relies on the extraction effectiveness, and developing a reliable extractor can be costly. A contrastive learning strategy, proposed by Zhuang et al. (2022), aims to maximize similarity between generated summaries and source documents while minimizing it between summaries and edited documents. Hosking et al. (2023) suggest an attributable opinion summarization system that encodes sentences as paths through a hierarchical discrete latent space, identifying common subpaths for an entity to generate the summary. Jin and Chen (2024) propose using a review from a review set as the hypothetic summary to carry out self-supervised summarization of product reviews.

3. Methods

Next, we introduce QGSumm, a novel framework for summarizing clinical notes, with a focus on capturing important patient-centered information in a self-supervised fashion. We propose a self-supervised domain adaptation strategy applied on the base model in Section 3.1. This strategy positive-contrastively learns from the original nursing notes, providing the summaries with an ability comparable to the original notes to resolve patient-related queries (Section 3.2). Using two augmentation blocks (Section 3.3), the model leverages patient metadata and temporal aspects. Assume a patient PT has a sequence of nursing notes $N = \{N_1, N_2, \dots, N_m\}$ sorted by time. Our objective is to obtain a summary S_i for note N_i from the distribution $P(S_i|N_i, PA, \{N_1, \dots, N_{i-1}\}, U)$, which is conditioned on the patient's metadata PA, information in the previous notes $\{N_1, \dots, N_{i-1}\}$, and the query U which helps guide the generation.

3.1. Base Model

The backbone of our framework is an off-the-shelf transformer-based language model with an encoder-decoder structure. We leverage its checkpoint M_0, fine-tuned for text summarization, as the **base model**, utilizing the pre-trained resources without training from scratch. However, M_0 exhibits limited clinical text comprehension. Hence, in **Stage 1** we improve the clinical text understanding of M_0's encoder, ENC_0, see Figure 2. Specifically, we employ ENC_0 from M_0 and the frozen decoder DEC_0 from the original pre-trained model, training ENC_0 by reconstructing N_i. This yields M with an improved encoder ENC which better grasps the complex semantics of nursing notes.

3.2. Training Objective

Since there is no ground truth summary available, in **Stage 2**, we adopt a self-supervised strategy for model M to generate high-quality, patient-

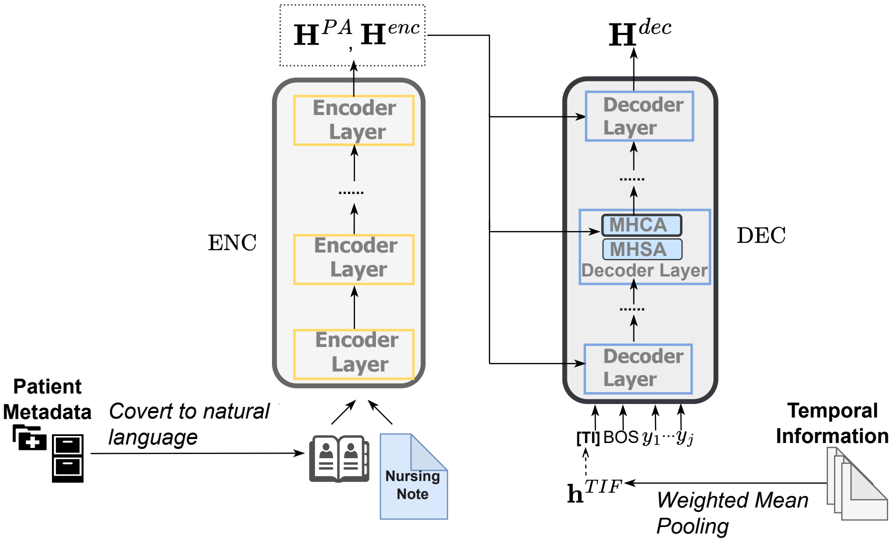


Figure 3: The proposed Temporal Information Fusion(TIF) block and the Patient Information Augmentation (PIA) block. The figure shows the process of deriving $\mathbf{H}^{dec}$ for generating the $(j+1)$th token in the summary.

centered summaries that can respond to patient-related queries effectively (Figure 2). We introduce a separate model R, a **query responder**, trained to handle patient-related queries such as predicting readmission risk using data available in the patient database. For example, given the note N_i or its summary S_i as input to R, the output will be the response to the corresponding query, like the readmission probability. As the training objective for M we minimize the discrepancy, calculated as the cross-entropy loss $\mathcal{L}_{CE}$, between the two responses based on either N_i or S_i respectively. This ensures that when responding to a certain patient-related query, using the summary will produce a response similar to that obtained using the nursing note, without ever showing the correct query answers to model M. To prevent M from generating summaries that are too verbose or direct "copy-paste" from the original notes, we introduce a length penalty term. The final loss function for a batch of nursing notes is:

$$\mathcal{L}_{summ} = \frac{1}{K}\sum_{i=1}^{K} \mathcal{L}_{CE}(\mathrm{R}(N_i), \mathrm{R}(S_i)) \times (1 + \lambda_1 e^{(\alpha - 0.5)}), \quad (1)$$

where

$$\alpha = \frac{\sum_{i=1}^{K} \mathrm{Len}(S_i)}{\sum_{i=1}^{K} \mathrm{Len}(N_i)}, \quad (2)$$

K is the batch size, and Len denotes the number of tokens in the document. The hyperparameter $\lambda_1 \in [0, 1]$ regulates the extent of the penalty, and is selected using the validation set, see details in C.2. Nondifferentiability from M and R is resolved using the straight-through Gumbel-softmax trick (Bengio et al., 2013; Jang et al., 2017).

3.3. Augmentation Blocks for the Context of the Patient

Temporal Information Fusion (TIF). A patient typically has sequential nursing notes to document the evolution of their condition. Therefore, the context provided by the prior notes helps understanding the patient's current status. We regard this as *temporal information* and include it during summarization to help the model to understand the patient's condition. For N_i, the embeddings of its previous notes are represented by the embeddings of their respective first tokens, which are special tokens indicating the start of each note. These embeddings are obtained at the last hidden state in the ENC, denoted as $\{\mathbf{h_1}, \mathbf{h_2}, \ldots, \mathbf{h_{i-1}}\}$, where $\mathbf{h_i} \in \mathbb{R}^d$ and d is the hidden state dimension. We aggregate past note representations via weighted mean pooling, emphasizing the most recent notes. In practice, initial weights $\beta_j, j = 1, \ldots, i-1$ for each past note N_i are based on the position in the sequence:

$\beta_1 = 1, \beta_2 = 2, \ldots, \beta_{i-1} = i - 1$. We use the normalized weights $\beta_j^{'} = \frac{\beta_j}{\beta_1+\beta_2+\ldots+\beta_{i-1}}$ to perform the weighted mean pooling:

$$\mathbf{h}^{TIF} = \text{MeanPooling}(\beta_1^{'}\mathbf{h}_1, \beta_2^{'}\mathbf{h}_2, ..., \beta_{i-1}^{'}\mathbf{h}_{i-1}), \quad (3)$$

where $\mathbf{h}^{TIF} \in \mathbb{R}^d$ represents the information fusion of the past notes. As shown in Figure 3, we prepend a special token `[TI]` at the beginning of the decoder input, representing temporal information with embedding $\mathbf{h}^{TIF}$. Consequently, the initial input to the decoder consists of `[[TI], [BOS]]`, where `[BOS]` indicating the start of generation. `[TI]` is replaced with the padding token `[PAD]` for nursing notes without past notes.

The model generates tokens auto-regressively, appending produced tokens to the decoder input when generating subsequent tokens. Thus, `[TI]` consistently prompts the model to focus on the patient's past context throughout the summarization process.

Patient Information Augmentation (PIA). We aim to obtain summaries focusing on the patient's condition by incorporating patient-level information into the model via cross-attention, facilitating the interaction of information on different levels. A patient's metadata PA typically comprises basic information, including age, gender, existing diagnoses, and performed procedures. We convert this metadata into patient information in natural language (one example in A.1), then encode it using ENC to derive patient embedding $\mathbf{H}^{PA} \in \mathbb{R}^{z \times d}$, where z represents the number of tokens in patient information. The source note is also encoded into embedding $\mathbf{H}^{enc} \in \mathbb{R}^{n \times d}$, where n denotes the number of tokens in the note given as input. On the decoder DEC side, let us assume the tokens input to the decoder at the current timestep are `[[TI], [BOS],` $y_1, \ldots, y_j$`]`. Consequently, the hidden representation passed to the lth decoder layer is $\mathbf{H}_l^{dec} \in \mathbb{R}^{(j+2) \times d}$. $\mathbf{H}_l^{dec}$ is updated using Multi-Head Self-Attention (MHSA) and Multi-Head Cross-Attention (MHCA) (Vaswani et al., 2017) involving $\mathbf{H}^{enc}$ and $\mathbf{H}^{PA}$. This facilitates the fusion of patient and note-level information:

$$\begin{aligned}\mathbf{H}_{l+1}^{dec} &= \text{MHCA}(\mathbf{H}^{enc}, \text{MHSA}(\mathbf{H}_l^{dec})) \\ &\quad + \lambda_2 \times \text{MHCA}(\mathbf{H}^{PA}, \text{MHSA}(\mathbf{H}_l^{dec})), \quad (4)\end{aligned}$$

$\lambda_2 \in [0, 1]$ is a hyperparameter to control the importance of patient meta information, selected using the validation set, see details in C.3. $\mathbf{H}_{l+1}^{dec}$ is the input to the next decoder layer, or if the lth layer is the final layer, it is the input to the language modeling head.

With these two augmentation blocks, we can obtain the final decoder state $\mathbf{H}^{dec} \in \mathbb{R}^{(j+2) \times d}$, which is input to the language modeling head for generating the $(j+1)$th token in the summary. The computation is shown in A.5

4. Experimental setup

4.1. Data

We use MIMIC-III (Johnson et al., 2016), an EHR database with clinical notes organized by admission. Notes are treated independently per admission due to the discontinuity between admissions of the same patient. We focus on the nursing notes within the clinical notes. After preprocessing (details in A.2), the numbers of notes in the training, validation and test sets are 149015, 10001, 3079 and the corresponding numbers of admissions are 13893, 1000, 1156.

4.2. Types of Queries

Two principles guide the query selection: (a) The query should be related to the patient and learnable by the query responder R; (b) Training data for R should be easily available. We propose two queries: **Readmission Prediction**, predicting if a patient will be readmitted within 30 days post-discharge, and **Phenotype Classification**, identifying which of 25 phenotypes the patient has (Harutyunyan et al., 2019). In the experiments, we combine these queries.

The query responder R is a classifier that predicts an answer to the query. Part of the training data is used to train R. When using R, we input the summary or the original note to predict the classification probabilities, and minimize the discrepancy between these predictions (see Section 3.2). As a baseline we include a query by minimizing the cosine similarity between the note and its summary. More details on queries and their combination are in A.3.

4.3. Detailed Settings

We use BART-Large-CNN[1] as the **base model**, a BART model (Lewis et al., 2020) fine-tuned for text summarization. For the **query responder**, we use Clinical-Longformer (Li et al., 2022), as it can handle long contexts, and fine-tune it on the selected queries. Hyperparameters for the models are specified in A.4.

Since our method is designed for scenarios where reference summaries are unavailable, we compare our

[1] https://huggingface.co/facebook/bart-large-cnn

method with **baselines** in zero-shot settings: BART-Large-CNN (*BART*) and Pegasus (Zhang et al., 2020a); and in 1-shot in-context learning prompting using GPT-4 (OpenAI, 2023) and BioMistral-7B (Labrak et al., 2024) (*BioMistral*). Consistent with QGSumm, we fine-tune the encoder of BART and Pegasus to reconstruct notes, enhancing their performance. We consider BART as the most relevant baseline for our method as it is the base model in our method, and hence represents performance without the proposed novel components.

For a fair comparison, we use six different prompts when evaluating GPT-4 and BioMistral: (1) the original prompt without additional information; (2) with patient information; (3) indicating the summary is for readmission prediction; (4) for phenotype classification; (5) for both readmission prediction and phenotype classification; and (6) including all previous notes as temporal information. The original prompt is shown in the main results due to negligible performance differences between prompts (1)-(5), and poorer performance with the prompt (6). A single summary example is included in the prompt to ensure that the generated summaries align with structural requirements. The content of prompts and full results for five different prompts are in C.5. Additional results with few-shot fine-tuning settings and extractive methods are in C.4, and more details about baselines in B.1.

4.4. Evaluation Metrics

Evaluating the quality of text summarization is challenging (Bhandari et al., 2020), especially in specialized fields and without reference summaries. Therefore, we employ multiple metrics to provide a comprehensive evaluation.

Automatic Evaluation Metrics. Metrics in the automatic evaluation are divided into three categories: 1) ***factuality and consistency***, 2) ***predictiveness***, and 3) ***conciseness***.

For consistency and factuality, we consider: (1) **UMLS-Recall**. We use QuickUMLS (Soldaini and Goharian, 2016) to extract Unified Medical Language System (UMLS) biomedical concepts from the nursing note and its summary. Recall is the proportion of concepts in the original note that are present in the summary. (2) **UMLS-FDR**. Denotes False Discovery Rate, quantifying the proportion of medical concepts in the summary that are not present in the original note. (3) **FactKB**. Evaluates factual consistency of summaries based on overall semantic information (Feng et al., 2023).

Predictiveness metric assesses whether the summary adequately contains patient key information, quantified as the ability to predict the patient's condition using the summary. Specifically, we conduct readmission prediction and phenotype classification using summaries from baselines and QGSumm. For a fair comparison, we use summaries from each method to fine-tune its respective classifier for the downstream task (similar to the query responder). This ensures the best evaluation result on predictiveness for each method, as illustrated in B.3. For readmission prediction, we report the weighted F1 and F1 of the positive class ("being readmitted"), and for phenotype classification, we report the F1-Macro.

Finally, we report the summary length as a percentage of the original note's length to assess conciseness. We do not enforce a strict maximum length for baselines as the model should be capable of determining the appropriate length autonomously.

Metrics used in the manual evaluation by a clinician. Without a reference summary, automatic evaluation metrics may not fully capture the quality of the summary. Therefore, we invite a clinician to manually evaluate the summaries generated from 25 nursing notes by 4 summarization method (25×4 summaries in total). The summaries are shown to the clinician in a randomized order and without showing the method that created them. Each summary is rated on four aspects from 1 to 5: (1) **Informativeness:** Whether the summary adequately captures essential information regarding the patient's condition; (2) **Fluency:** Whether the summary is well-written and easy to understand. (3) **Consistency:** How well the summary aligns with the nursing note in factuality. (4) **Relevance:** Whether the summary is concise and not contain unnecessary information. Detailed grading criteria are in B.2.

5. Results and Discussion

5.1. Automatic evaluation

Conciseness, Consistency, and Factuality. There is a trade-off between UMLS-Recall and summary length (conciseness). As shown in Table 1, our method balance medical information consistency (measured by UMLS-Recall) and conciseness. GPT-4 captures more medical information but is less concise,

Table 1: Results of automatic evaluation. Lower values are better for Length and UMLS-FDR, higher values for the other metrics. The subscripts denote standard deviation. “Orig. Notes” means using original nursing notes as such for readmission and phenotype prediction. “Re+Ph” means using “Readmission Prediction and Phenotype Classification” as the query. Results from **best** and <u>2nd best</u> method under each metric are bolded and underlined.

Type	Method	Consistency and Factuality			Conciseness	Predictiveness		
		UMLS-Recall	UMLS-FDR	FactKB	Length	Readmission		Phenotype
						Weighted F1	F1	Macro F1
	Orig. Notes	-	-	-	-	$85.2_{0.5}$	$19.7_{1.9}$	$28.7_{0.5}$
Baselines	BART	$36.4_{9.0}$	$\mathbf{8.70}_{6.2}$	$0.78_{0.16}$	**31.9%**	$78.8_{0.4}$	$11.1_{0.9}$	$20.5_{0.3}$
	GPT-4	$\mathbf{59.2}_{8.3}$	$44.2_{7.6}$	$0.77_{0.17}$	53.6%	$\mathbf{85.6}_{0.6}$	$\mathbf{21.5}_{2.0}$	$\underline{23.6}_{0.6}$
	BioMistral	$55.4_{9.9}$	$50.0_{8.7}$	$0.68_{0.14}$	69.2%	$80.1_{0.6}$	$10.7_{1.3}$	$21.4_{0.4}$
	Pegasus	$32.2_{8.7}$	$47.4_{7.9}$	$0.65_{0.12}$	51.9%	$77.1_{0.4}$	$9.3_{0.8}$	$17.5_{0.5}$
QGSumm	-Similarity	$53.1_{7.2}$	$\underline{20.7}_{6.7}$	$\mathbf{0.82}_{0.13}$	51.7%	$79.5_{0.6}$	$12.0_{1.2}$	$22.4_{0.4}$
	-Re+Ph	$\underline{58.8}_{7.9}$	$24.1_{6.4}$	$\underline{0.80}_{0.14}$	<u>48.2%</u>	$\underline{84.2}_{0.5}$	$\underline{17.2}_{1.6}$	$\mathbf{25.1}_{0.5}$

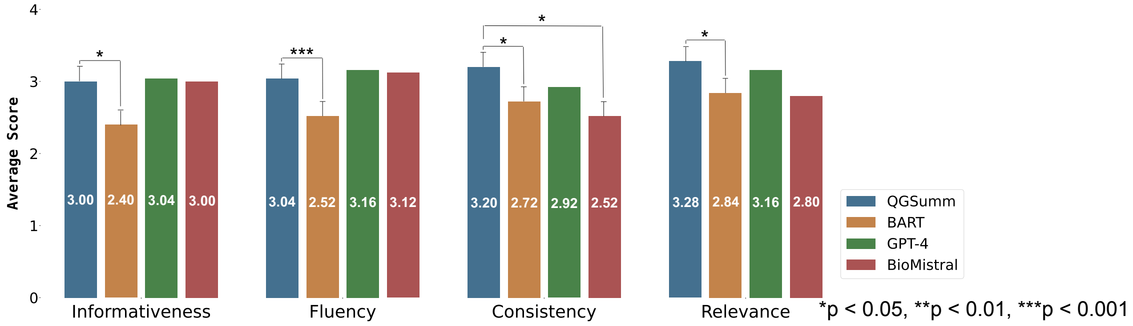


Figure 4: Results of the manual evaluation by a clinician. The average scores in four metrics are reported for QGSumm, BART, GPT-4, and BioMistral. “*” denotes the result of the significance test, calculated using a two-tailed Binomial test on the pairwise win-rates, i.e., we count the number of notes where QGSumm has a score higher or lower than a comparison method and test for the null hypothesis that the win-rate is 0.5.

whereas BART produces concise but less informative summaries. Similar to GPT-4, BioMistral tends to produce summaries that are not concise. Summaries from BART maintain high levels in factuality (UMLS-FDR and FactKB). Our method also has strong performance on relevant metrics. We find that although LLMs, such as GPT-4 and BioMistral, excel in language understanding, they do not perform well on factuality. One possible reason is their tendency to rephrase or expand the original notes, potentially introducing inconsistent information, i.e. hallucination, which can cause a relatively high UMLS-FDR.

Predictiveness. Results are shown in Table 1. In the readmission prediction task, our method closely follows GPT-4, surpassing all other baselines. We see that our method outperforms BART significantly in weighted F1 score (84.2 vs. 78.8) and F1 score of the positive class (17.2 vs. 11.1). This shows the effectiveness of the adaptation strategy guiding the model with useful queries. Interestingly, we find that using the summary from GPT-4 for this task outperforms using the original notes, and our method's summaries perform nearly as well, highlighting their quality. In phenotype classification, our method performs the best, outperforming BART in Macro F1 (25.1 vs. 20.5). Even when using similarity alone, it surpasses BART (22.4 vs. 20.5). Notably, our method's training objective does not involve true labels of the queries, and the query responder remains frozen during training, preventing query data leakage into M. Instead, the strong predictiveness of our summaries follows from emphasizing relevant patient-related information present in the original notes. Although specialized in text summarization, Pegasus has weak performance on all predictiveness metrics.

5.2. Manual Evaluation by a Clinician

To avoid excessive manual work, we select 3 baselines: BART, GPT-4, and BioMistral in addition to QGSumm. BART was selected as the main baseline, as it is the base model in our method but without the novel components. GPT-4 and BioMistral are selected due to their strong performance in the automatic evaluation. Average scores for each method across four metrics are shown in Figure 4. A case study comparing example summaries from these methods can be found in C.1.

QGSumm vs BART. QGSumm significantly outperforms BART on all four metrics. This indicates that our domain adaptation strategy enables the model to generate higher-quality summaries from the medical personnel's perspective, containing refined and important patient information with fewer hallucinations and increased readability. Despite producing longer summaries, our method achieves a higher relevance score from the clinician, suggesting the base model struggles to identify key information and focuses on unnecessary details. Our model can effectively enhance this aspect.

QGSumm vs GPT-4 and BioMistral. GPT-4 and BioMistral perform similarly to QGSumm in Informativeness but excel slightly in Fluency by rephrasing and clarifying abbreviations. However, the rephrasing can introduce factual inconsistencies, and the tendency to infer additional content may reduce factuality. An example of this is provided in the case study in C.1. Consequently, QGSumm scores higher in Consistency, which is essential in the clinical setting. Furthermore, it generates more concise summaries with higher scores in Relevance. However, due to the small sample size, the only statistically significant difference in these comparisons was the improvement of QGSumm compared to Biomistral in consistency, and further work is needed for more conclusive results.

5.3. Ablation Study

We conducted an ablation study to: **(1)** evaluate the performance of the model without the proposed augmentation blocks: without the Patient Information Augmentation (w/o PIA), without the Temporal Information Fusion (w/o TIF), and without both blocks (w/o PIA+TIF); and **(2)** evaluate the effectiveness of the query guidance by using "Readmission Prediction" and "Phenotype Classification" as separate queries, instead of the combined query used earlier.

Table 2 shows that the removal of augmentation blocks causes weighted F1 and macro F1 scores decrease in all settings, indicating that both blocks enhance the predictiveness of the summary. Removing TIF results in a larger score drop in F1 scores, underlining the importance of temporal information to understand the patient's current status. Conversely, the removal of PIA degrades the performance on factuality (UMLS-FDR and FactKB) more than the removal of TIF, while the removal of TIF affects more the UMLS-Recall, i.e., fewer medical concepts are captured. On the other hand, employing different queries allows the model to focus on different aspects of the original note, resulting in summaries varying across different metrics. While there are some exceptions, the general trend is that the combined query yields overall better performance than using either query alone. Full results and a more detailed analysis are presented in C.6.

5.4. Discussion and Conclusion

User need -oriented summarization. A high-quality summary should facilitate efficient understanding of the relevant content by clinical personnel, especially for a nursing note the summary should capture the patient's condition. Our method employs patient-related queries, indirectly ensuring that the summary centers around the patient's status. The summaries generated with different queries can be seen as coming from distinct conditional distributions and parts of the semantic space, allowing control over content and granularity. This facilitates a more flexible and user need -oriented summary generation. We show in C.6 that the queries can guide summaries to focus on specific aspects of the original note. For instance, broad queries produce comprehensive summaries, while detailed queries yield focused summaries on specific conditions like cardiovascular health.

Design choices for information augmentation. One challenge is how to efficiently integrate information into the model without excessive computational cost. We use cross-attention to allow the patient's metadata to efficiently interact on multiple levels during summary generation. In contrast, for temporal information in previous notes, using cross-attention in a similar manner might make it difficult for the model to balance attention across the current

Table 2: Results of the ablation study. We show the change in the value of the metric after removing different augmentation blocks/using different queries. ↓ denotes a decrease in the score and ↑ denotes an increase.

	Weighted F1	Macro F1	UMLS-Recall	UMLS-FDR	FactKB
QGSumm(-Re+Ph)	84.2	25.1	58.8	24.1	0.80
w/o PIA	↓ 2.6	↓ 1.4	↓ 1.4	↑ 4.7	↓ 0.04
w/o TIF	↓ 4.1	↓ 1.6	↓ 3.8	↑ 1.9	↓ 0.01
w/o PIA+TIF	↓ 4.8	↓ 2.3	↓ 4.4	↑ 5.6	↓ 0.04
-Readmission	↓ 1.8	↓ 1.2	↓ 0.6	↓ 1.4	↓ 0.02
-Phenotype	↓ 2.3	↑ 0.5	↓ 0.3	↑ 12	↓ 0.01

note, past notes, and patient information, and increase computational challenges with long sequences. Hence, we represent the temporal information, obtained by weighted mean pooling from previous note representations, as the first token of the decoder's input. This strategy is intuitive, as information from previous notes naturally precedes the summary of the current note.

Interpretation of the evaluation metrics. The metrics used in the automatic evaluation have limitations as they do not conclusively reflect the quality of the summary, and come with trade-offs. For example, a good performance in predictiveness and medical information consistency (UMLS-Recall) may not be due to the high quality of the summary but rather caused by copying the source note, resulting in a lack of conciseness and fluency. Conversely, as the summary becomes more concise, it may become less informative. Furthermore, models used to measure factuality have inherent biases. They are potentially weak at recognizing patient-related information due to the dissimilarity between their training domain and clinical data. We attempt to mitigate the impact of these limitations by comprehensively considering multiple metrics, and including the manual evaluation by a clinician. However, there remains a need for more conclusive evaluation metrics, particularly when ground truth is unavailable.

Limitations. (1) Our current approach produces summaries of individual nursing notes, and lack the long context and support for multiple note summarization. (2) There is room for more exploration on the formulation of the clinical queries, such as generative queries. Also, when investigating the combined effects of multiple queries, further exploration using multi-task learning methods could be beneficial. (3) Due to the workload, the number of summaries assessed in the manual evaluation is limited to 100.

Conclusion. We presented a novel self-supervised nursing note summarization method, where the main innovation was the introduction of query guidance, which successfully directed the summaries to include desired content. In the manual evaluation by a professional clinician, our method significantly outperformed a specialized open text summarization model, BART-Large-CNN, in all metrics. Of the other baselines, the proprietary GPT-4 had the closest performance to our method and was better than the other baselines. In the automatic evaluation, our method outperformed GPT-4 in factual consistency, having fewer hallucinated facts without sacrificing the correct content. The same trend was seen in the manual evaluation as higher average consistency for our method. Hence, our approach can produce more reliable summaries, clearing obstacles for responsible clinical use of LLMs. Our method demonstrates the feasibility of domain adaptation for pre-trained text summarization models without explicit supervision, and the effectiveness of self-supervised strategies to guide conditional summarization to specific interests.

Acknowledgments

This work was supported by the Research Council of Finland (Flagship programme: Finnish Center for Artificial Intelligence FCAI, and grants 352986, 358246) and EU (H2020 grant 101016775 and NextGenerationEU).

References

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, 2023.

Griffin Adams, Han-Chin Shing, Qing Sun, Christopher Winestock, Kathleen Mckeown, and Noémie Elhadad. Learning to revise references for faithful summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4009–4027, 2022.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, 2020.

Eric Chu and Peter Liu. Meansum: A neural model for unsupervised multi-document abstractive summarization. In *International Conference on Machine Learning*, pages 1223–1232. PMLR, 2019.

Martina A Clarke, Jeffery L Belden, Richelle J Koopman, Linsey M Steege, Joi L Moore, Shannon M Canfield, and Min S Kim. Information needs and information-seeking behaviour analysis of primary care physicians and nurses: a literature review. *Health Information & Libraries Journal*, 30 (3):178–190, 2013.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36, 2024.

Hady Elsahar, Maximin Coavoux, Jos Rozen, and Matthias Gallé. Self-supervised and controlled multi-document opinion summarization. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1646–1662, 2021.

Shangbin Feng, Vidhisha Balachandran, Yuyang Bai, and Yulia Tsvetkov. Factkb: Generalizable factuality evaluation using language models enhanced with factual knowledge. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 933–952, 2023.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, Dongfang Xu, Matthew MM Churpek, and Majid Afshar. Summarizing patients' problems from hospital progress notes using pre-trained sequence-to-sequence models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2979–2991, 2022.

Yanjun Gao, Dmitriy Dligach, Timothy Miller, and Majid Afshar. Overview of the problem list summarization (probsum) 2023 shared task on summarizing patients' active diagnoses and problems from electronic health record progress notes. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 461–467, 2023.

Amanda Hall and Graham Walton. Information overload within the health care system: a literature review. *Health Information & Libraries Journal*, 21 (2):102–108, 2004.

Hrayr Harutyunyan, Hrant Khachatrian, David C Kale, Greg Ver Steeg, and Aram Galstyan. Multitask learning and benchmarking with clinical time series data. *Scientific Data*, 6(1):96, 2019.

Tom Hosking, Hao Tang, and Mirella Lapata. Attributable and scalable opinion summarization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8488–8505, 2023.

Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. Clinicalbert: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparametrization with gumble-softmax. In *International Conference on Learning Representations (ICLR 2017)*, 2017.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7b. *arXiv preprint arXiv:2310.06825*, 2023.

Lingyun Jin and Jingqiang Chen. Self-supervised opinion summarization with multi-modal knowledge graph. *Journal of Intelligent Information Systems*, 62(1):191–208, 2024.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Liwei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific Data*, 3 (1):1–9, 2016.

Neel Kanwal and Giuseppe Rizzo. Attention-based clinical note summarization. In *Proceedings of the 37th ACM/SIGAPP Symposium on Applied Computing*, pages 813–820, 2022.

Wenjun Ke, Jinhua Gao, Huawei Shen, and Xueqi Cheng. Consistsum: Unsupervised opinion summarization with the consistency of aspect, sentiment and semantic. In *Proceedings of the fifteenth ACM International Conference on Web Search and Aata Mining*, pages 467–475, 2022.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kundan Krishna, Sopan Khosla, Jeffrey P Bigham, and Zachary C Lipton. 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 (Volume 1: Long Papers)*, pages 4958–4972, 2021.

Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. Biomistral: A collection of open-source pretrained large language models for medical domains. *arXiv preprint arXiv:2402.10373*, 2024.

Andrew Lampinen, Ishita Dasgupta, Stephanie Chan, Kory Mathewson, Mh Tessler, Antonia Creswell, James McClelland, Jane Wang, and Felix Hill. Can language models learn from explanations in context? In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 537–563, 2022.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, 2020.

Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. Clinical-longformer and clinical-bigbird: Transformers for long clinical sequences. *arXiv preprint arXiv:2201.11838*, 2022.

Fenglin Liu, Bang Yang, Chenyu You, Xian Wu, Shen Ge, Zhangdaihong Liu, Xu Sun, Yang Yang, and David Clifton. Retrieve, reason, and refine: Generating accurate and faithful patient instructions. *Advances in Neural Information Processing Systems*, 35:18864–18877, 2022a.

Puyuan Liu, Chenyang Huang, and Lili Mou. Learning non-autoregressive models from search for unsupervised sentence summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7916–7929, 2022b.

Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. *IEEE Transactions on Knowledge and Data Engineering*, 35(1):857–876, 2021.

Rada Mihalcea and Paul Tarau. Textrank: Bringing order into text. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 404–411, 2004.

Hans Moen, Laura-Maria Peltonen, Juho Heimonen, Antti Airola, Tapio Pahikkala, Tapio Salakoski, and Sanna Salanterä. Comparison of automatic summarisation methods for clinical free text notes. *Artificial Intelligence in Medicine*, 67:25–37, 2016.

Mizuho Nishio, Takaaki Matsunaga, Hidetoshi Matsuo, Munenobu Nogami, Yasuhisa Kurata, Koji Fujimoto, Osamu Sugiyama, Toshiaki Akashi, Shigeki Aoki, and Takamichi Murakami. Fully automatic summarization of radiology reports using natural language processing with large language models. *Informatics in Medicine Unlocked*, 46: 101465, 2024.

R OpenAI. Gpt-4 technical report. *ArXiv*, 2303, 2023.

Heather C O'Donnell, Rainu Kaushal, Yolanda Barrón, Mark A Callahan, Ronald D Adelman, and Eugenia L Siegler. Physicians' attitudes towards copy and pasting in electronic note writing. *Journal of General Internal Medicine*, 24:63–68, 2009.

Rimma Pivovarov and Noémie Elhadad. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association : JAMIA*, 22:938 – 947, 2015.

Akseli Reunamo, Laura-Maria Peltonen, Reetta Mustonen, Minttu Saari, Tapio Salakoski, Sanna Salanterä, and Hans Moen. Text classification model explainability for keyword extraction–towards keyword-based summarization of nursing care episodes. In *MEDINFO 2021: One World, One Health–Global Partnership for Digital Innovation*, pages 632–636. IOS Press, 2022.

Thomas Searle, Zina Ibrahim, James Teo, and Richard JB Dobson. Discharge summary hospital course summarisation of in patient electronic health record text with clinical concept guided deep pre-trained transformer models. *Journal of Biomedical Informatics*, 141:104358, 2023.

Han-Chin Shing, Chaitanya Shivade, Nima Pourdamghani, Feng Nan, Philip Resnik, Douglas Oard, and Parminder Bhatia. Towards clinical encounter summarization: Learning to compose discharge summaries from prior notes. *arXiv preprint arXiv:2104.13498*, 2021.

Luca Soldaini and Nazli Goharian. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR Workshop, SIGIR*, pages 1–4, 2016.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867, 2020.

Matthew Tang, Priyanka Gandhi, Md. Ahsanul Kabir, Christopher Zou, Jordyn Blakey, and Xiao Luo. Progress notes classification and keyword extraction using attention-based deep learning models with bert. *ArXiv*, abs/1910.05786, 2019.

Eva Törnvall and Susan Wilhelmsson. Nursing documentation for communicating and evaluating care. *Journal of Clinical Nursing*, 17(16):2116–2124, 2008.

Dave Van Veen, Cara Van Uden, Maayane Attias, Anuj Pareek, Christian Bluethgen, Malgorzata Polacin, Wah Chiu, Jean-Benoit Delbrouck, Juan Zambrano Chaves, Curtis Langlotz, et al. Radadapt: Radiology report summarization via lightweight domain adaptation of large language models. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 449–460, 2023.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, pages 1–9, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 2017.

Mengqian Wang, Manhua Wang, Fei Yu, Yue Yang, Jennifer Walker, and Javed Mostafa. A systematic review of automatic text summarization for biomedical literature and ehrs. *Journal of the American Medical Informatics Association*, 28(10): 2287–2297, 2021.

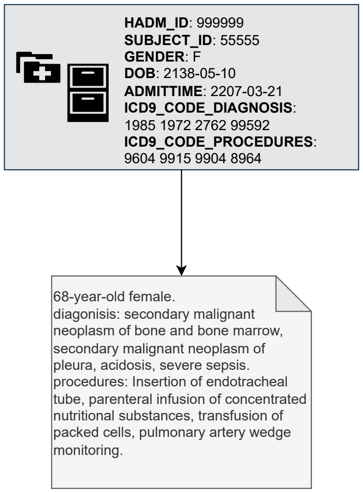
Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR, 2020a.

Longxiang Zhang, Renato Negrinho, Arindam Ghosh, Vasudevan Jagannathan, Hamid Reza Hassanzadeh, Thomas Schaaf, and Matthew R Gormley. Leveraging pretrained models for automatic summarization of doctor-patient conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3693–3712, 2021.

Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D Manning, and Curtis Langlotz. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5108–5120, 2020b.

Ming Zhong, Yang Liu, Suyu Ge, Yuning Mao, Yizhu Jiao, Xingxing Zhang, Yichong Xu, Chenguang Zhu, Michael Zeng, and Jiawei Han. Unsupervised multi-granularity summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4980–4995, 2022.

Haojie Zhuang, Wei Emma Zhang, Jian Yang, Congbo Ma, Yutong Qu, and Quan Z Sheng. Learning



HADM_ID: 999999
SUBJECT_ID: 55555
GENDER: F
DOB: 2138-05-10
ADMITTIME: 2207-03-21
ICD9_CODE_DIAGNOSIS:
1985 1972 2762 99592
ICD9_CODE_PROCEDURES:
9604 9915 9904 8964

68-year-old female.
diagonisis: secondary malignant neoplasm of bone and bone marrow, secondary malignant neoplasm of pleura, acidosis, severe sepsis.
procedures: Insertion of endotracheal tube, parenteral infusion of concentrated nutritional substances, transfusion of packed cells, pulmonary artery wedge monitoring.

Figure 5: Convert artificial patient metadata to a natural language description.

from the source document: Unsupervised abstractive summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4194–4205, 2022.

Appendix A. Additional Details for Implementation

A.1 Patient Metadata

Figure 5 shows one artificial example of how patient information is obtained in natural language from structural metadata. In the MIMIC-III database, we retrieve patient identifiers (“SUBJECT_ID”), gender information(“GENDER”), and date of birth (“DOB”) from the “PATIENTS” table. Information regarding admission identifiers (“HADM_ID”) and admission time (“ADMITTIME”) are obtained from the “ADMISSIONS” table, while diagnosis codes and procedure codes are sourced from “DIAGNOSES_ICD” and “PROCEDURES_ICD” tables, respectively.

A.2 Data Preprocessing

Following prior research (Harutyunyan et al., 2019; Huang et al., 2019), we perform filtering on admission records and nursing notes. Initially, we filter out specific admission cases: (1) cases of in-hospital mortality and admissions categorized as ”NEWBORN”; (2) cases containing diagnosis codes outside HCUP CCS code groups. We retain only admissions containing clinical notes categorized as ”Nursing/other”. After the initial filtering, we get the statistics of The distribution of the number of nursing notes per admission, shown in Table 3. Subsequently, we apply a length limit to nursing notes, filtering out those with more than 800 tokens or fewer than 50 tokens. Finally, we filtered out admission cases with more than 100 nursing notes. Nursing notes in these cases typically represent out-of-distribution information or are irrelevant to the care episode.

Table 3: The distribution of the number of nursing notes per admission.

percentiles	number of nursing notes per admission
25%	3
50%	6
75%	13
90%	35
100%	913

We preprocess nursing notes following (Huang et al., 2019). In addition, we expand certain frequently occurring abbreviations found multiple times in each note, such as “pt” (patient), “cv” (cardiovascular), and “resp” (respiratory), to aid the model’s understanding of the notes. By random sampling, we collect 10001 nursing notes from 1000 admissions as the validation set. For the test set, we randomly select 1516 admissions and sample 3079 nursing notes from these admissions. We only use 3079 notes for testing due to the cost of the use of GPT-4. The nursing notes in remaining admissions are included in the training set.

A.3 Details of Implementation for the Query Responder

In addition to Readmission Prediction and Phenotype Classification, we also employ another query, Contrastive Next Note Prediction. Hence, totally five different queries are used in the experiment.

Contrastive Next Note Prediction. Given a nursing note pair (N, N'), we regard the query about whether N' is the successor note of N as a prediction of the **patient’s future status**. To train the query responder R for the next note prediction, we create two note pairs for each nursing note, where the positive pair (N, N_{pos}) comprises the note and its successor in the sequence, and the negative pair (N, N_{neg}) contains the note and a randomly chosen non-consecutive note. If N is the patient's last nursing note, we use the patient's discharge summary and a random note from other patients to construct the positive and negative pairs. The query is formulated as binary classification, and the output of R is the probability of each pair being the positive pair containing the consecutive notes:

$$p_{pos} = \mathrm{R}(N, N_{pos}), \quad p_{neg} = \mathrm{R}(N, N_{neg}), \tag{5}$$

$$p'_{pos} = \mathrm{R}(S, N_{pos}), \quad p'_{neg} = \mathrm{R}(S, N_{neg}). \tag{6}$$

The learning objective in this case is:

$$\min \mathcal{L}_{CE}([p_{pos}, p_{neg}], [p'_{pos}, p'_{neg}]). \tag{7}$$

Readmission Prediction. Readmission information is easily retrieved from the hospital's database and is closely related to the patient. The readmission prediction query is formulated as a 2-class classification task to predict whether the patient will be readmitted within 30 days of discharge, which reflects the **patient's future condition**.

The result of the readmission prediction is in the form of $[p_{pos}, p_{neg}]$, indicating the probability of "being readmitted" and "not being readmitted". The learning objective is the same as the Equation 7.

Phenotype Classification. Classifying a patient's diagnosis status or phenotype is a query to the **patient's current status**. Following Harutyunyan et al. (2019), phenotype classification is formulated as a multi-label classification, where ICD-9 diagnosis codes mapped by HCUP CCS code groups[2] are categorized into 25 classes. Therefore, the responder outputs the probability distribution of the phenotype as $[p_1, p_2, ..., p_{25}]$.

The prediction process can be formulated as:

$$[p_1, p_2, ..., p_{25}] = \mathrm{R}(N), \quad [p'_1, p'_2, ..., p'_{25}] = \mathrm{R}(S). \tag{8}$$

Consequently, the learning objective is:

$$\min \mathcal{L}_{CE}([p_1, p_2, ..., p_{25}], [p'_1, p'_2, ..., p'_{25}]). \tag{9}$$

Readmission Prediction and Phenotype Classification. We investigate the combined utilization of two queries, readmission prediction, and phenotype classification, to see if joint guidance is more effective. After obtaining the result of readmission prediction $[p_1^r, p_2^r]$ and the result of phenotype classification $[p_1^c, \ldots, p_{25}^c]$, we integrate them by converting the results into a 50-class probability distribution.

Table 4: The number of nursing notes are used for training, validation and testing.

Query	Training	Validation	Testing
Next Note Prediction	100000	5000	17458
Readmission Prediction	35000	10001	17458
Phenotype Classification	149015	10001	17458

Training of R**.** We use nursing notes in the training set to train the query responder R. The data statistics are presented in Table 4.

To address the class imbalance issue in the readmission prediction task, we conduct oversampling for notes in the positive class ("being readmitted") and undersampling for notes in the negative class ("not being readmitted"). This results in 35000 nursing notes being used for training.

A.4 Hyperparameter Setting

We present hyperparameter settings of QGSumm and the query responder R in Table 5. The configuration of hyperparameters for the base model's architecture keeps the same as the original configuration of BART-Large-CNN[3]. We set the maximum length of the summary to 500 tokens, allowing for flexibility as we aim for the model to autonomously determine the appropriate length. We use Adam optimizer (Kingma and Ba, 2014) to optimize the model. The experiments are ran on one Tesla A100 with 80G memory and one Tesla P100 with 16G memory.

A.5 Next Token Prediction

As illustrated in 3.3, we obtain the hidden representation $\mathbf{H}^{dec}$ from the decoder's final layer. The process of generating the next token then can be abstracted as:

$$[\mathbf{H}^{enc}, \mathbf{H}^{PA}] = \mathrm{ENC}(N_i, PA), \tag{10}$$

$$\mathbf{H}^{dec} = \mathrm{DEC}(\mathbf{H}^{enc}, \mathbf{H}^{PA}, [[TI], [BOS], y_1, \ldots, y_j]), \tag{11}$$

$$\mathbf{v} = \mathrm{LMH}(\mathbf{H}^{dec}), \tag{12}$$

$$\mathbf{v}' = \text{ST-GumbelSoftmax}(\mathbf{v}). \tag{13}$$

LMH (Language Modeling Head) maps $\mathbf{H}^{dec}$ to a probability vector $\mathbf{v} \in \mathbb{R}^{vs}$ over the vocabulary of size vs. $\mathbf{v}$ is processed using the straight-though gumbel

[2] https://hcup-us.ahrq.gov/toolssoftware/ccs/ccsfactsheet.jsp

[3] https://huggingface.co/facebook/bart-large-cnn/blob/main/config.json

Table 5: Details of the hyperparameter setting.

	Hyperparameter	Choices
QGSumm	learning rate	{1e-5, 2e-5, 5e-5, 2e-4, 5e-4}
	number of training epochs	3
	number of training epochs for note reconstruction	1
	λ_1	{0.1, 0.3, 0.5, 0.7, 0.9, 1.0}
	λ_2	{0.0, 0.1, 0.3, 0.6, 0.8, 1.0}
	decoder layers being augmented by PIA	{all 12 layers, first 6 layers, last 6 layers}
R	learning rate	{2e-5, 5e-5, 2e-4, 5e-4, 1e-3}
	number of training epochs for next note prediction	2
	number of training epochs for readmission prediction	2
	number of training epochs for phenotype classification	3

softmax trick, denoted as ST-GumbelSoftmax, resulting in a one-hot vector $\mathbf{v}' \in \mathbb{R}^{vs}$ providing the index of the $(j+1)$th token.

Appendix B. Additional Details for Baselines and Evaluation

B.1 Baselines

Choice of the baselines

BART-Large-CNN (Lewis et al., 2020): It is chosen as the base model for its excellent performance on text summarization as well as less computation cost than its peers. We consider it as one baseline in the experiment to illustrate performance without the proposed novel components. **Pegasus** (Zhang et al., 2020a): It is a transformer-based pre-trained model specialized in abstractive summarization, widely recognized as a baseline model in many studies on text summarization. **BioMistral-7B** (Labrak et al., 2024): It is an open-source instruction-based LLM adapted from Mistral (Jiang et al., 2023) for the medical domain. It achieves state-of-the-art performance in supervised fine-tuning benchmarks compared to other open-source medical language models. **GPT-4** (OpenAI, 2023): It is a proprietary LLM representing state-of-the-art on general NLP task. We employ one-shot in-context learning to prompt GPT-4 and BioMistral-7B, providing one summary example to ensure the structure of generated summaries aligns with the requirements. We show prompt examples in C.5. Consistent with our method's settings, the maximum length of the summary is set to 500 tokens.

In Table 6, we provide the statistics on the total number of parameters for QGSumm and baselines.

We also consider two representative extractive methods, TextRank (Mihalcea and Tarau, 2004) and Lead-40%, and report the evaluation results of them as a reference. We do not compare our methond and baselines with them, since extractive methods are out of the scope of this work. In TextRank, we utilize MPNet (Song et al., 2020) to obtain sentence embeddings. In Lead-40%, we use the first 40% of the content of the note as a summary.

Table 6: The statistics on the total number of parameters for QGSumm and baselines (except GPT-4, whose parameter count remains undisclosed).

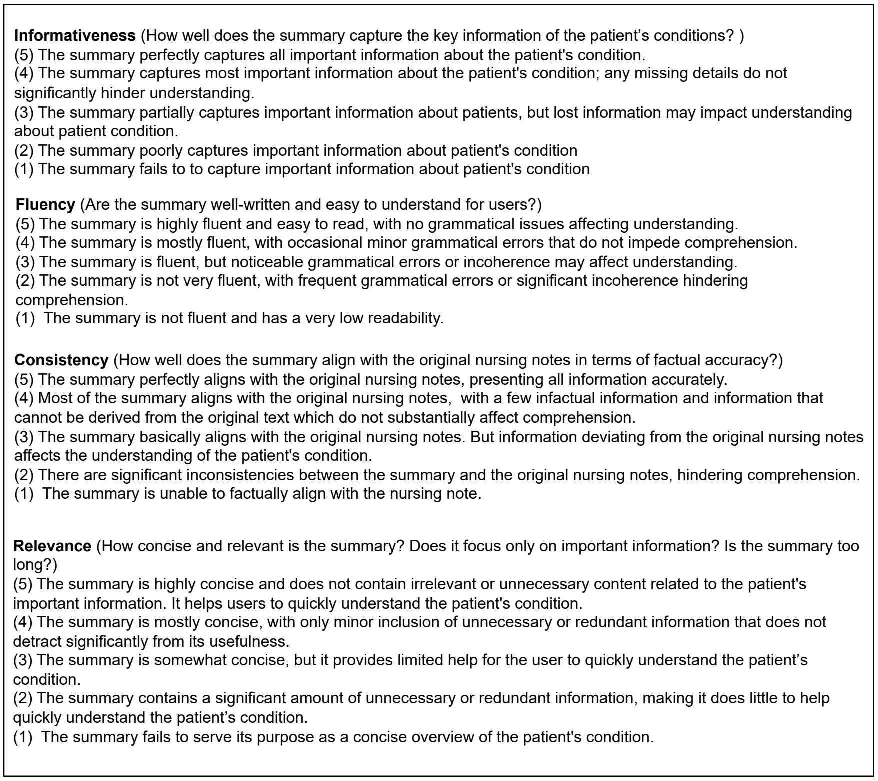
	number of parameters
QGSumm	457M
BART-Large-CNN	406M
Pegasus	568M
BioMistral	7B

Few-shot Adaptation

In addition, we show the performance of baselines in few-shot fine-tuning settings for reference. We randomly sample 10 nursing notes from the training set and pair them with their corresponding summaries generated by GPT-4 as the sliver ground truth to create the training data for 10-shot fine-tuning of BART-Large-CNN, BioMistral-7B, and Pegasus. The training data is transformed into instructions-formatted prompts for fine-tuning BioMistral-7B. We fine-tune BART-Large-CNN and Pegasus for 8 and 9 epochs, respectively, with a learning rate of 0.0005. As for BioMistral-7B, we fine-tune it using QLoRA adaptation for 7 epochs with a learning rate set to 0.0002.

B.2 Grading Criteria of Manuel Evaluation

Figure 6 shows the detailed grading criteria we provide to the clinician for manual evaluation. The score for each metric ranges from 1 to 5.



Informativeness (How well does the summary capture the key information of the patient's conditions?)
(5) The summary perfectly captures all important information about the patient's condition.
(4) The summary captures most important information about the patient's condition; any missing details do not significantly hinder understanding.
(3) The summary partially captures important information about patients, but lost information may impact understanding about patient condition.
(2) The summary poorly captures important information about patient's condition
(1) The summary fails to to capture important information about patient's condition

Fluency (Are the summary well-written and easy to understand for users?)
(5) The summary is highly fluent and easy to read, with no grammatical issues affecting understanding.
(4) The summary is mostly fluent, with occasional minor grammatical errors that do not impede comprehension.
(3) The summary is fluent, but noticeable grammatical errors or incoherence may affect understanding.
(2) The summary is not very fluent, with frequent grammatical errors or significant incoherence hindering comprehension.
(1) The summary is not fluent and has a very low readability.

Consistency (How well does the summary align with the original nursing notes in terms of factual accuracy?)
(5) The summary perfectly aligns with the original nursing notes, presenting all information accurately.
(4) Most of the summary aligns with the original nursing notes, with a few infactual information and information that cannot be derived from the original text which do not substantially affect comprehension.
(3) The summary basically aligns with the original nursing notes. But information deviating from the original nursing notes affects the understanding of the patient's condition.
(2) There are significant inconsistencies between the summary and the original nursing notes, hindering comprehension.
(1) The summary is unable to factually align with the nursing note.

Relevance (How concise and relevant is the summary? Does it focus only on important information? Is the summary too long?)
(5) The summary is highly concise and does not contain irrelevant or unnecessary content related to the patient's important information. It helps users to quickly understand the patient's condition.
(4) The summary is mostly concise, with only minor inclusion of unnecessary or redundant information that does not detract significantly from its usefulness.
(3) The summary is somewhat concise, but it provides limited help for the user to quickly understand the patient's condition.
(2) The summary contains a significant amount of unnecessary or redundant information, making it does little to help quickly understand the patient's condition.
(1) The summary fails to serve its purpose as a concise overview of the patient's condition.

Figure 6: The grading criteria provided to the clinician for manual evaluation.

B.3 Evaluations on Predictiveness

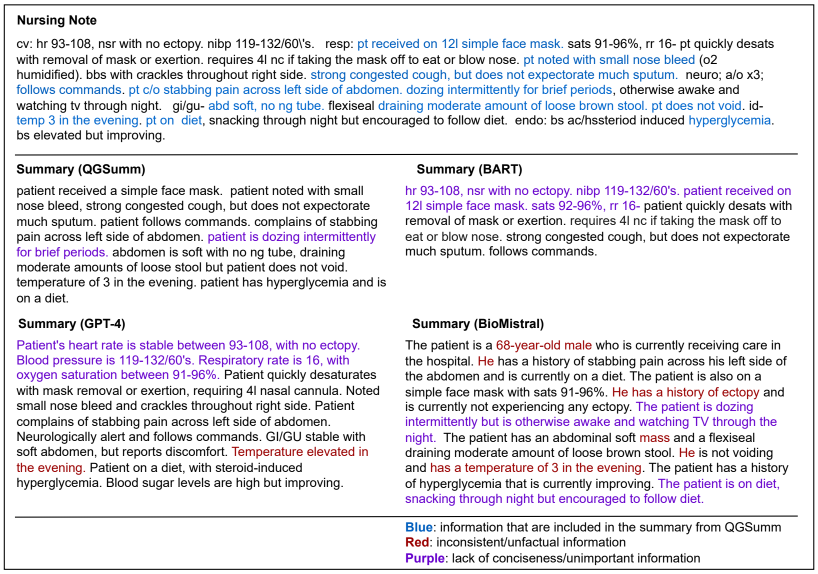
Based on the assumption that a high-quality summary, by effectively capturing critical patient information, can better predict both the current and future status of patients, we introduce predictiveness as a novel metric to evaluate the quality of the summaries.

Specifically, we assess predictiveness by measuring the performance of readmission prediction and phenotype classification using summaries generated by various baselines. The evaluation workflow for this metric is as follows:

- Generate summaries of notes in the test set for each method (baselines and QGSumm);
- For summaries from one method, split them into 10 folds;
- Fine-tune the **own** query responder of each method using its summaries, and compute the metrics through 10-fold cross-validation.

For instance, to evaluate the predictiveness of summaries from BART, we begin with the trained query responder R as described in A.3. We fine-tune R using the BART-generated summaries of nursing notes from the test set and then compute the relevant metrics through 10-fold cross-validation. Consequently, we obtain multiple responders tailored to each method, specifically for readmission prediction and phenotype classification.

We fine-tune each method's query responder using its own summaries to **ensure the most accurate evaluation of predictiveness** for each method. For comparison, we also experimented with using only GPT-4's summaries to fine-tune a single query responder across all methods. As shown in Table 7, the results got worse for all methods distinct from GPT-4, which might be due to the dissimilarity of the semantic spaces between summaries from GPT-4 and summaries from others. By optimizing a separate predictor for each model we get the best possible predictions for each one, leading to a fair comparison.



Nursing Note

cv: hr 93-108, nsr with no ectopy. nibp 119-132/60\'s. resp: pt received on 12l simple face mask. sats 91-96%, rr 16- pt quickly desats with removal of mask or exertion. requires 4l nc if taking the mask off to eat or blow nose. pt noted with small nose bleed (o2 humidified). bbs with crackles throughout right side. strong congested cough, but does not expectorate much sputum. neuro; a/o x3; follows commands. pt c/o stabbing pain across left side of abdomen. dozing intermittently for brief periods, otherwise awake and watching tv through night. gi/gu- abd soft, no ng tube. flexiseal draining moderate amount of loose brown stool. pt does not void. id- temp 3 in the evening. pt on diet, snacking through night but encouraged to follow diet. endo: bs ac/hssteriod induced hyperglycemia. bs elevated but improving.

Summary (QGSumm)

patient received a simple face mask. patient noted with small nose bleed, strong congested cough, but does not expectorate much sputum. patient follows commands. complains of stabbing pain across left side of abdomen. patient is dozing intermittently for brief periods. abdomen is soft with no ng tube, draining moderate amounts of loose stool but patient does not void. temperature of 3 in the evening. patient has hyperglycemia and is on a diet.

Summary (BART)

hr 93-108, nsr with no ectopy. nibp 119-132/60's. patient received on 12l simple face mask. sats 92-96%, rr 16- patient quickly desats with removal of mask or exertion. requires 4l nc if taking the mask off to eat or blow nose. strong congested cough, but does not expectorate much sputum. follows commands.

Summary (GPT-4)

Patient's heart rate is stable between 93-108, with no ectopy. Blood pressure is 119-132/60's. Respiratory rate is 16, with oxygen saturation between 91-96%. Patient quickly desaturates with mask removal or exertion, requiring 4l nasal cannula. Noted small nose bleed and crackles throughout right side. Patient complains of stabbing pain across left side of abdomen. Neurologically alert and follows commands. GI/GU stable with soft abdomen, but reports discomfort. Temperature elevated in the evening. Patient on a diet, with steroid-induced hyperglycemia. Blood sugar levels are high but improving.

Summary (BioMistral)

The patient is a 68-year-old male who is currently receiving care in the hospital. He has a history of stabbing pain across his left side of the abdomen and is currently on a diet. The patient is also on a simple face mask with sats 91-96%. He has a history of ectopy and is currently not experiencing any ectopy. The patient is dozing intermittently but is otherwise awake and watching TV through the night. The patient has an abdominal soft mass and a flexiseal draining moderate amount of loose brown stool. He is not voiding and has a temperature of 3 in the evening. The patient has a history of hyperglycemia that is currently improving. The patient is on diet, snacking through night but encouraged to follow diet.

Blue: information that are included in the summary from QGSumm
Red: inconsistent/unfactual information
Purple: lack of conciseness/unimportant information

Figure 7: One artificial nursing note and its summaries from QGSumm, BART, GPT-4, and BioMistral, respectively.

Table 7: Reults on predictiveness get worse when using only summaries from GPT-4 to finetune the query responder. Weighted F1 and Macro F1 are reported for readmission prediction and phenotype classification, respectively.

	Weighted F1	Macro F1
QGSumm	↓1.6	↓0.7
BART	↓1.9	↓0.9
BioMistral	↓0.9	↓0.4
Pegasus	↓1.7	↓0.7

Appendix C. Additional Experimental Results

C.1 Case Study

One artificial nursing note and its corresponding summaries generated by QGSumm, BART, GPT-4, and BioMistral are presented in Figure 7. In the original nursing note, the content highlighted in blue indicates information included in the summary generated by our approach. We can see that our approach captures most of the important patient information. However, some details, such as cardiovascular and respiratory conditions, are overlooked. The summary from BART only covers information from the first half of the nursing note, suggesting the limitation in understanding long context. Summaries from GPT-4 and BioMistral contain more patient information but lack conciseness. These models achieve fluency by rephrasing notes and expanding abbreviations. However, BioMistral struggles with maintaining factuality, often excessively reasoning about the patient's personal information and condition.

C.2 Effects of the Length Penalty Coefficient

The effects of varying the length penalty coefficient are shown in Figure 8(a). When λ_1 increases, the generated summaries become more concise. However, once λ_1 exceeds 0.5, there is a notable decrease in medical information consistency, accompanied by a decline in predictiveness performance. One potential explanation for this phenomenon is that within the range of 0.1 to 0.5, λ_1 facilitates the refinement of nursing notes by filtering out unnecessary information. However, surpassing 0.5 in the value of λ_1 results in a stricter penalty, which causes the omission of the patient key information for obtaining more concise summaries. To strike a balance between conciseness and consistency, we ultimately set λ_1 to 0.5.

C.3 Effects of the Importance of Patient Meta Information

λ_2 regulates the contribution of patient metadata to the summarization process. The incorporation of patient meta information helps maintain the factuality

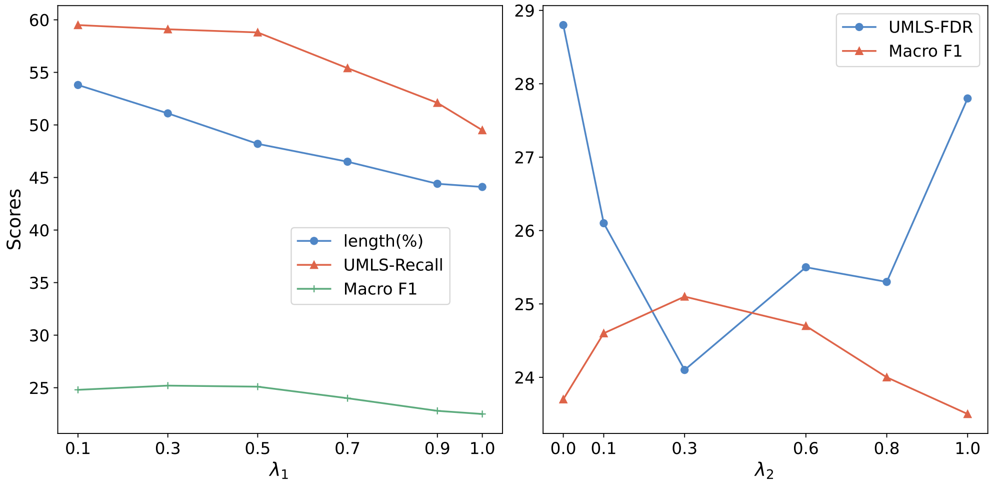


Figure 8: Effects of various values of λ_1 (Equation 1) and λ_2 (Equation 4). (a): λ_1 affects both the length and the medical information consistency of the generated summary, resulting in changes in the summary's predictiveness; (b): λ_2 regulates the importance of the PIA block. The PIA is applied to all decoder layers.

Table 8: Additional results of: (1) automatic evaluation on few-shot fine-tuning and extractive methods; (2) using five different queries. Lower values are better for Length and UMLS-FDR, higher values for the other metrics. The subscripts denote standard deviation.

Type	Method	Consistency and Factuality			Conciseness	Predictiveness		
						Readmission		Phenotype
		UMLS-Recall	UMLS-FDR	FactKB	Length	Weighted F1	F1	Macro F1
Few-Shot	BART	$52.5_{7.3}$	$44.5_{7.1}$	$0.76_{0.15}$	65.0%	$82.2_{0.5}$	$14.4_{1.3}$	$21.1_{0.4}$
	BioMistral	$57.2_{10.2}$	$49.1_{7.8}$	$0.70_{0.15}$	68.8%	$81.7_{0.4}$	$10.2_{1.1}$	$22.0_{0.4}$
	Pegasus	$35.1_{8.4}$	$52.6_{7.7}$	$0.70_{0.18}$	57.4%	$80.5_{0.8}$	$12.5_{1.8}$	$18.3_{0.6}$
Extractive	Lead-40%	$42.7_{6.7}$	$0.30_{2.6}$	$0.99_{0.06}$	40.0%	$83.1_{0.6}$	$12.6_{1.5}$	$21.7_{0.5}$
	TextRank	$58.5_{7.9}$	$0.08_{1.4}$	$0.95_{0.12}$	51.9%	$81.9_{0.7}$	$14.4_{1.7}$	$23.3_{0.5}$
QGSumm	-Similarity	$53.1_{7.2}$	$20.7_{6.7}$	$0.82_{0.13}$	51.7%	$79.5_{0.6}$	$12.0_{1.2}$	$22.4_{0.4}$
	-NextNote	$56.4_{8.0}$	$35.2_{7.1}$	$0.77_{0.11}$	49.3%	$80.8_{0.6}$	$11.7_{1.4}$	$23.2_{0.6}$
	-Readmission	$58.2_{7.5}$	$22.7_{6.5}$	$0.78_{0.14}$	46.2%	$82.4_{0.5}$	$18.2_{1.6}$	$23.9_{0.5}$
	-Phenotype	$58.5_{7.4}$	$36.2_{6.9}$	$0.79_{0.13}$	48.0%	$81.9_{0.5}$	$13.4_{1.5}$	$25.6_{0.6}$
	-Re+Ph	$58.8_{7.9}$	$24.1_{6.4}$	$0.80_{0.14}$	48.2%	$84.2_{0.5}$	$17.2_{1.6}$	$25.1_{0.5}$

of the summary. However, as shown in Figure 8(b), this influence is not consistently beneficial, with the optimal effect observed when λ_2 is set to 0.3. Additionally, excessively large λ_2 causes the model to prioritize patient metadata over the content of the nursing note, which degrades the quality of the generated summary, reflected as reduced predictiveness. Therefore, we set λ_2 to 0.3 for our method.

C.4 Results on Few-shot and Extractive Methods

We present additional results on few-shot fine-tuning settings and extractive methods, as shown in Table 8.

After fine-tuning with 10 summaries generated by GPT-4, BART and Pegasus produce summaries that include more medical and patient-related information, as indicated by higher F1-score and UMLS-Recall. However, their performance declines in certain factual metrics and their summaries become less concise, mirroring the performance trend seen with GPT-4. BioMistral's performance has not changed much after few-shot adaptation. Considering the inherent bias from the limitations of GPT-4, we do not compare QGSumm with few-shot fine-tuning methods in the main text.

C.5 Prompts for GPT-4 and BioMistral

Examples of the prompts used for evaluating GPT-4 and BioMistral are shown in Figure 9. We observe that when a summary example is not included in the prompt, the model sometimes generates outputs that fail to meet the structural requirements, such as producing lists of sentences. This issue is more prevalent when using GPT-4.

Table 9: We conduct evaluations on 100 nursing notes using six different prompts for GPT-4 and BioMistral. Lower values are better for Length and UMLS-FDR, higher values for the other metrics.

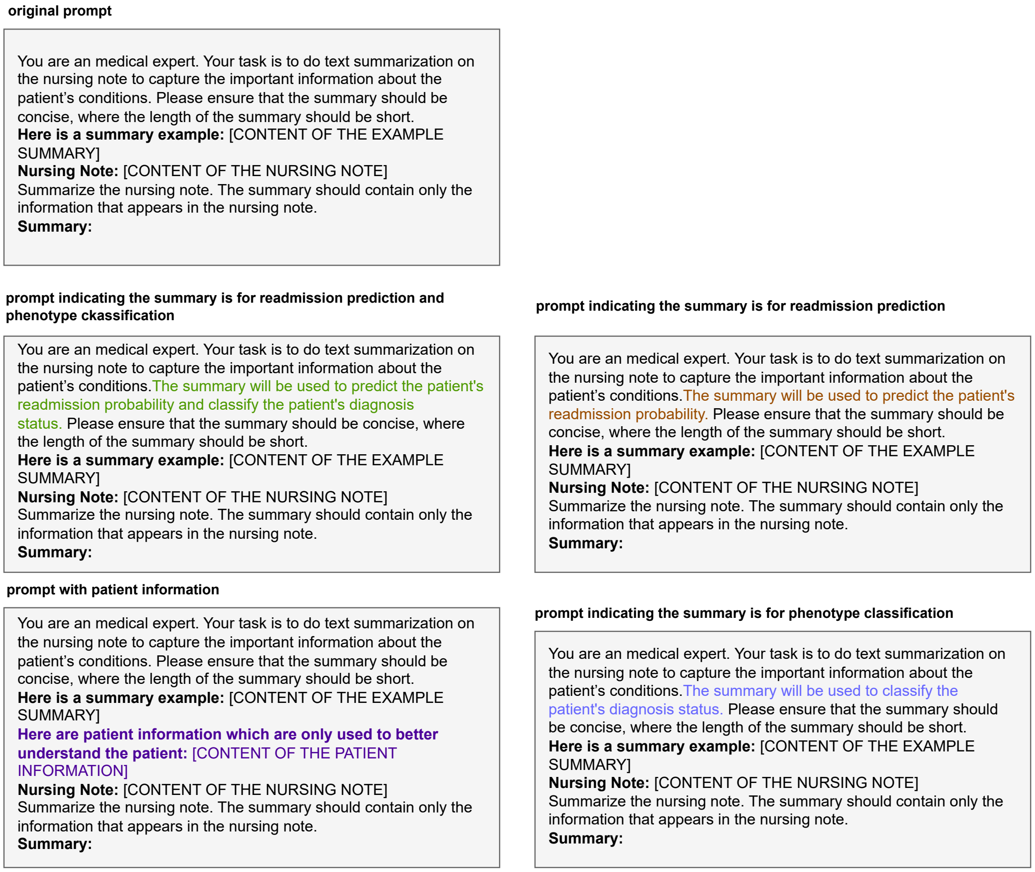
		UMLS-Recall	UMLS-FDR	FactKB	Length	Weighted F1	Macro F1
GPT-4	Original Prompt	56.1	44.9	0.73	56.5%	84.6	23.4
	- Patient Information	53.8	48.0	0.76	63.6%	83.2	21.7
	- Readmission	55.6	44.4	0.64	54.5%	83.5	24.0
	- Penotype	54.4	45.1	0.61	54.9%	83.5	21.1
	- Re+Ph	56.1	45.0	0.64	55.7%	84.6	22.6
	- Temporal Info	52.2	50.2	0.65	60.3%	82.4	21.5
BioMistral	Original Prompt	53.8	50.3	0.69	67.7%	79.7	21.2
	- Patient Information	55.1	54.4	0.64	70.1%	78.6	22.0
	- Readmission	53.1	50.3	0.63	67.1%	78.2	20.2
	- Penotype	52.2	51.2	0.63	65.9%	77.6	20.9
	- Re+Ph	53.8	51.7	0.61	66.7%	79.3	21.4
	- Temporal Info	50.4	58.6	0.62	68.2%	76.0	19.3

To determine the optimal prompt content, we conduct evaluations using six different prompts, with the results shown in Table 9. Given the cost of GPT-4, we limit the evaluation to 100 randomly selected nursing notes from the test set. Compared to the original prompt, adding patient information or queries did not result in significant differences in performance for either GPT-4 or BioMistral. When considering temporal information by including all previous notes in the prompt, we observed that including temporal information interfered with the summarization of the current note, causing both GPT-4 and BioMistral to miss key information from the current note. Consequently, we choose to use the original prompts in the main experiments, as they yielded more balanced results across different metrics.

Furthermore, we note that unlike our method, using prompts that explicitly mention readmission prediction or phenotype classification does not improve performance on those specific tasks. This highlights the effectiveness of using the query to parametrically guide the model's behavior.

C.6 Effectiveness of the Query Guidance

We provide full results of employing five different queries shown in Table 8. We can observe: (1) Regarding predictiveness, employing queries closely related to patients and focusing on readmission and phenotype information yield superior performance compared to other queries. As expected, the method with phenotype-related queries performs the best in phenotype classification, while the method with readmission-related queries is the best in readmission prediction. This highlights the effectiveness of guiding the summarization with queries, and **different queries enable the summary to concentrate on different aspects of the original note.** (2) Using similarity as guidance can produce summaries that are more similar to the original notes, resulting in higher scores on general consistency and factuality. However, summaries generated under this configuration tend to be longer and often sacrifice predictiveness and informativeness regarding medical concepts, **demonstrating the limitations of the unconstrained guidance signal.** (3) Using Next Note Prediction as the query results in weaker performance compared to patient-related queries, indicating that **queries lacking meaningful context fail to guide the model effectively.** (4) When employing joint guidance with both readmission and phenotype information, our method consistently achieves excellent performance across all metrics. This indicates that **combining different guidance signals can help in producing better summaries,** and further research is needed to explore this aspect in depth.



original prompt

You are an medical expert. Your task is to do text summarization on the nursing note to capture the important information about the patient's conditions. Please ensure that the summary should be concise, where the length of the summary should be short.
Here is a summary example: [CONTENT OF THE EXAMPLE SUMMARY]
Nursing Note: [CONTENT OF THE NURSING NOTE]
Summarize the nursing note. The summary should contain only the information that appears in the nursing note.
Summary:

prompt indicating the summary is for readmission prediction and phenotype ckassification

You are an medical expert. Your task is to do text summarization on the nursing note to capture the important information about the patient's conditions.The summary will be used to predict the patient's readmission probability and classify the patient's diagnosis status. Please ensure that the summary should be concise, where the length of the summary should be short.
Here is a summary example: [CONTENT OF THE EXAMPLE SUMMARY]
Nursing Note: [CONTENT OF THE NURSING NOTE]
Summarize the nursing note. The summary should contain only the information that appears in the nursing note.
Summary:

prompt indicating the summary is for readmission prediction

You are an medical expert. Your task is to do text summarization on the nursing note to capture the important information about the patient's conditions.The summary will be used to predict the patient's readmission probability. Please ensure that the summary should be concise, where the length of the summary should be short.
Here is a summary example: [CONTENT OF THE EXAMPLE SUMMARY]
Nursing Note: [CONTENT OF THE NURSING NOTE]
Summarize the nursing note. The summary should contain only the information that appears in the nursing note.
Summary:

prompt with patient information

You are an medical expert. Your task is to do text summarization on the nursing note to capture the important information about the patient's conditions. Please ensure that the summary should be concise, where the length of the summary should be short.
Here is a summary example: [CONTENT OF THE EXAMPLE SUMMARY]
Here are patient information which are only used to better understand the patient: [CONTENT OF THE PATIENT INFORMATION]
Nursing Note: [CONTENT OF THE NURSING NOTE]
Summarize the nursing note. The summary should contain only the information that appears in the nursing note.
Summary:

prompt indicating the summary is for phenotype classification

You are an medical expert. Your task is to do text summarization on the nursing note to capture the important information about the patient's conditions.The summary will be used to classify the patient's diagnosis status. Please ensure that the summary should be concise, where the length of the summary should be short.
Here is a summary example: [CONTENT OF THE EXAMPLE SUMMARY]
Nursing Note: [CONTENT OF THE NURSING NOTE]
Summarize the nursing note. The summary should contain only the information that appears in the nursing note.
Summary:

Figure 9: We use five different prompts for evaluating GPT-4 and BioMistral.