Conditioning Clinical Text Generation for User Control

Anonymous ACL submission

Abstract

Deploying Natural Language Generation (NLG) systems in clinical settings poses a significant challenge despite the recent advancements in Large Language Models (LLMs), as they remain prone to hallucinations and factual errors, requiring human oversight to prevent harmful outcomes in practice. In this paper, we propose methods to condition LLMs for controlled clinical text generation, allowing clinicians to influence the writing style and select topics to dynamically refine the output. We further demonstrate, on the basis of the BioNLP ACL'24 Shared Task 'Discharge Me!', how models conditioned for user control can be trained and evaluated on existing datasets. Our results show an improvement of more than 9% over previous submissions with our baseline model, and more than 34% with user control, highlighting the significant impact of clinician-guided generation on relevance, accuracy, and factual consistency. Our work also reveals a flaw in traditional datasets—namely, that stylistic variations in the output cannot be derived from the input—and underscore the need for augmenting traditional datasets.

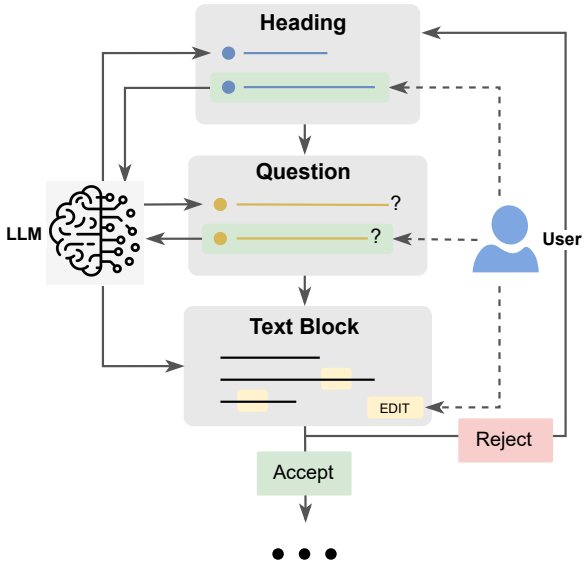

Figure 1: The interactive workflow enabled by our topic-level generation control. The user generates clinical documents with the depicted workflow topic-by-topic, sequentially accepting or submitting suggestions for headings and questions, and editing, accepting or rejecting text blocks.

1 Introduction

Large language models (LLMs) like OpenAI's GPTs (OpenAI et al., 2024; Brown et al., 2020), Google's PaLM (Anil et al., 2023) and Gemini (Team et al., 2024), and lately Meta's Llama (Touvron et al., 2023a,b; Dubey et al., 2024) have shown remarkable versatility across a wide range of applications, including healthcare (Singhal et al., 2023; Huang et al., 2024). In clinical environments, these models offer opportunities to automate tasks such as summarizing clinical notes, supporting diagnostic decisions, and streamlining patient communication (Hirosawa et al., 2023; Soleimani et al., 2024; Ruinelli et al., 2024; Liu et al., 2023; Patel and Lam, 2023; Van Veen et al., 2024; Zaretsky et al., 2024; Were et al., 2010) . However, the deployment of AI in critical domains like healthcare faces significant challenges, since the cost of mistakes is significantly higher compared to most other fields, as even minor inaccuracies in clinical contexts might lead to serious consequences, such as misdiagnoses, incorrect treatments or other harmful patient outcomes.

One of the biggest concerns with AI-generated content in clinical settings is their tendency for hallucinations and factual errors (Ji et al., 2023; Lin et al., 2024; Tang et al., 2023; Dada et al., 2024). For this reason, AI-generated content cannot be accepted at face value and must always be verified by human experts before being incorporated into practice. Ethical considerations further complicate this process, calling for human clinicians to hold accountability for medical decisions through rig-

orous oversight over AI generated content (Meskó and Topol, 2023; Omiye et al., 2024).

Yet, the process of verifying AI-generated content is itself burdensome. Clinicians already face substantial cognitive workloads, and manually reviewing generated documents—especially when consolidating information from various sources—can negate the potential efficiency gains promised by automation. This introduces a paradox: while AI has the potential to alleviate some of the documentation and synthesis burdens, it risks adding new cognitive demands by requiring continuous oversight.

Addressing this paradox is essential for realizing the full potential of AI in clinical settings without increasing risks or workloads. To strike this balance, AI systems must be designed to offer control and transparency to human users. Clinicians need tools to shape the generation process, ensuring that outputs are accurate and aligned with clinical contexts, communication styles, and guidelines. In clinical settings, where accuracy and accountability are critical, such control over AI-output is a necessity (Meskó and Topol, 2023).

Our research aims to address this need by proposing strategies for conditioning clinical text generation to provide users with more granular control over content selection and writing style. Specifically, we introduce a system that naturally breaks down the generation process into distinct, manageable writing subtasks, reducing the complexity of both content generation and human verification while moving away from traditional black-box models that limit user involvement. By empowering users to influence the generation process dynamically at the subtask level, we aim to bridge the gap between automation and human oversight in clinical workflows.

However, traditional datasets do not adequately support this kind of user control, despite containing target texts with significant variations in writing style that can't be inferred from input data, and models being penalized for deviating from the target writing style regardless of task performance. To tackle this, we augment such datasets with authoring guidelines and topic segmentations, allowing models to be conditioned on both style and content, enabling clinicians to impose their preferences during generation.

Through our approach, we demonstrate that explicitly conditioning clinical text generation for user control over content and style significantly enhances model performance across multiple metrics, indicating improvements in the relevance, accuracy, and factual consistency of the generated output.

Our key contributions[1] are as follows:

- We set a new baseline on the BioNLP ACL'24 Shared Task 'Discharge Me!' leaderboard with simpler methods and less training than previous submissions.
- We propose methods to augment traditional datasets for enabling user control over both the content and style of clinical text generation.
- We develop strategies for interactive, topic-specific generation that enables granular user control and show empirically their importance with automated evaluation methods.

2 Related Work

In recent years, there has been increasing research into LLM-based clinical text generation, highlighting its potential in automating the production of discharge summaries (Ando et al., 2022; Ellershaw et al., 2024; Clough et al., 2024; Dubinski et al., 2024), brief hospital courses (BHC) (Hartman and Campion, 2022; Hartman et al., 2023; Searle et al., 2023), and radiology reports (Alfarghaly et al., 2021; Wang et al., 2023; Yang et al., 2023). Some studies even suggest that physicians often prefer AI-generated clinical texts over manually written ones (Van Veen et al., 2024). However, most of these approaches have treated clinical text generation as an end-to-end generation task, without offering opportunities for user intervention and control.

A recent example of this approach can be seen in the BioNLP ACL'24 Shared Task "Discharge Me!" (Xu et al., 2024) which focuses on the automated generation of discharge summaries. However, the complexity of clinical texts, which often require external sources of information and are influenced by the personal writing styles of different clinicians, makes an end-to-end approach less feasible in practice. These challenges point to the need for a more flexible generation process that allows users to control specific aspects of the output, such as content selection and style.

Controlled text generation (CTG) is a growing area of research aimed at providing users with more influence over the generated content. This involves

[1] All source code will be released upon paper acceptance.

integrating specific control conditions, such as enforcing a professional tone or ensuring the use of domain-specific terminology, while maintaining fluency, coherence, and relevance in the generated text (Zhang et al., 2023). Prior work in this area has explored different control mechanisms that can be directly applied to clinical text generation, such as structure control (Yang and Klein, 2021; Zou et al., 2021), general style control (Keskar et al., 2019), and personal style control (Tao et al., 2024).

Moreover, recent studies have explored using Question-Answer (QA) pairs as a blueprint to guide the text generation process. This approach has been shown to reduce hallucinations and improve the factual consistency of generated content (Narayan et al., 2023; Huot et al., 2023). It is based on the Question Under Discussion (QUD) theory (Roberts, 2012), which states that all utterances in a discourse (Van Kuppevelt, 1995) serve to answer either implicit or explicit questions. Building on these insights, we adapt the QUD framework for clinical document generation. In this context, we treat clinical documents as responses to a series of implicit questions that naturally arise from the document's intended purpose. These questions are typically addressed in a structured or patterned manner throughout the document, even in cases where the document appears unstructured, lacking consistent sections or headings.

By applying fine-grained topic segmentation, we try to reveal this underlying structure. We generate a sequence of headings and QUDs, with corresponding text segments acting as the answers to these questions. This approach allows us to align individual topics with distinct writing subtasks, enabling users to exert control over a simplified generation process while maintaining the structured nature expected in clinical documentation.

3 Conditioning Text Generation for User Control

We explore two strategies to adapt Large Language Models (LLMs) for conditioned clinical text generation, intended to provide intuitive control over the generated output: topic-level structured generation and authoring guidelines. These strategies, however, unveil certain limitations inherent to traditional datasets, particularly in the context of clinical text generation, that need to be addressed.

Traditionally, training datasets are built on the assumption that increasing the number of training examples naturally enhances model performance, with the idea that more data leads to better generalization. However, our experience indicates that existing datasets may not always capture the full range of information necessary for a model to perform well. For example, a wide range of medical documents exhibit significant variations in quality, format, and writing styles—even when addressing the same task (Horwitz et al., 2013; Edwards et al., 2014; Hultman et al., 2019). Experienced clinicians can even often recognize their colleagues in clinical practice by the distinct style of the written reports. These stylistic variations, however, cannot always be inferred from the input data alone. For instance, different parts of a discharge summary are authored by different individuals, each with their own distinct writing habits and preferences.

It is necessary to close this information gap in traditional datasets for multiple reasons: (a) many metrics used for evaluation are sensitive to these stylistic variations, as evidenced by this work, and (b) since the required stylistic variations cannot be inferred from the input alone, performance metrics are afflicted by randomness and do not faithfully reflect overall task performance.

To hold models accountable for stylistic deviations of generated output, we therefore extend the the task definition. Rather than simply generating documents, we require models to author documents according to specific guidelines, which are passed alongside the input. Conditioning the text generation to adhere to specific authoring guidelines also enables users to provide their own individual guidelines during inference.

Furthermore, typical summarization and text generation datasets are designed for end-to-end tasks, where all output is produced at once from the input data. This makes it difficult to build systems that allow for user intervention and control during the generation process. To overcome this, we segment texts into distinct topics, decomposing the overall generation task into smaller, more manageable subtasks. Each topic is then addressed through (a) topic headings, which serve as intuitive and concise summaries of the key content or themes within a text block, and (b) questions, which articulate the purpose of the text block and connect it to the document's overall intent. This conditioning of models enables structured, topic-level generation that is both controllable and intervenable by users (cf. Fig. 1).

3.1 Augmenting Datasets for User Control

We utilize the *Discharge Me!* challenge[2], part of the BioNLP ACL'24 Shared Tasks, for training and evaluating our models due to its clinical relevance and challenging nature. Additionally, its leaderboard provides a strong baseline. The task focuses on automating the generation of hospital course summaries and discharge instructions, traditionally time-intensive tasks for clinicians.

The dataset consists of 109,168 discharge summaries (DS) from the MIMIC-IV dataset, each containing a Brief Hospital Course (BHC) and a Discharge Instructions (DI) section. The BHC section is typically found in the middle of the discharge summary, following details on patient history and treatments during the current visit. The DI section is generally located at the end of the note. Additionally, each discharge summary is linked to at least one radiology report and typically one ICD chief complaint, along with multiple ICD codes.

To address the aforementioned limitations, we augment the *Discharge Me!* dataset by (a) segmenting each target text into topics and (b) generating individual authoring guidelines for each target text. We employ *Llama 3.1 70B Instruct* for this task, as LLMs have shown to effectively replace human annotators (Gilardi et al., 2023; Perez et al., 2022).

Topic Segmentation

We task an LLM with fine-grained topic segmentation of the target texts (such as the BHC or DI sections of the discharge summaries). The LLM is prompted to segment each target text t_i into smaller text blocks t_i^k. Simultaneously, it generates topic-specific headings h_i^k and questions q_i^k that capture the meaning and relevance of each segment for the intended audience, as well as the Question Under Discussion (QUD) for the segment. For instance, given the target text t_i, the LLM is instructed to return the segmentation as:

$$seg(t_i) = (h_i^1, q_i^1, t_i^1, ..., h_i^n, q_i^n, t_i^n).$$

Afterwards, we perform a post-processing step to align the segmented text blocks $t_i^1, \ldots, t_i^n$ back to the original text t_i. This is done using diff methods that identify differences at the word level (defined as whitespace-delimited character sequences). Successive differences are grouped into blocks, and only those segmentations that have at most single-word differences are matched back to the original text. These differences encompass minor corrections of misspellings or formatting. We then replace the generated text segments t_i^j with the matching text segments in the original text to retain the original formatting as much as possible. This applies to 93.61% of the segmentations of discharge instructions and 81.15% of the segmentations of brief hospital course reports.

Authoring Guidelines

We explore two types of authoring guidelines: style guidelines and writing instructions. In both cases, the LLM prompts are carefully engineered to avoid answer leakage by instructing the LLM to not use terms or phrases from the source text, to not quote or give examples from the patient records, and not to reveal patient-specific details. This yields very generic authoring guidelines that do not include any patient-specific information.

Style Guidelines describe stylistic features of the target text, including tone, document format, layout, composition, text structure, and language use (including abbreviations and medical jargon).

Writing Instructions provide explicit instructions on how to write the target text. Beyond the stylistic features above, they also cover aspects such as the purpose and intent of the text, the intended audience, and the text outline. The main difference, however, lies in the instructive wording.

3.2 Instruction Tuning for Conditioned Text Generation

While it is possible to prompt instruction-tuned LLMs to follow authoring guidelines or to generate structured output topic-by-topic, prompt engineering has it's limitations. Therefore, we specifically fine-tune all our models with instruction-tuning for conditioned text generation as follows.

Let $C = \{\texttt{none}, \texttt{topics}\}$ denote the possible configurations for structuring the generation output for control and $G = \{\texttt{none}, \texttt{styles}, \texttt{instructions}\}$ denote the possible configurations for using authoring guidelines. Then each prompt $prompt_i(c, g)$ with $c \in C, g \in G$ contains a user message $user_i(c, g)$ and assistant message $assistant_i(c)$ tailored to the configuration with explicit user instructions for generation.

The **user message** $user_i(c, g)$ contains the clinical context, that we get by concatenating the discharge summary s_i with the associated radiology reports $r_i^1, \ldots, r_i^j$. To condition the model on authoring guidelines, we append the respective guide-

[2] https://stanford-aimi.github.io/discharge-me

lines $g_i(t)$ to the clinical context. The model is instructed to generate text that complies with these guidelines. For topic-level structured generation, $user_i(g, c)$ provides explicit instructions on how to generate the structured input, depending on the choice $c \in C$.

The **assistant message** $assistant_i(c)$ contains the target text $t_i \in \{di_i, bhc_i\}$. To condition the model for topic-level structured generation, we prepend the topic-specific headings h_i^k and questions q_i^k to the text blocks $t_i^1, \ldots, t_i^n$, which are parsed from the topic segmentation $seg(t_i)$. Each element is wrapped in respective HTML tags. This allows us to stop and continue generation as needed using HTML tags as control sequences to detect the different stages in the process, allowing users to manipulate intermediate results (cf. Fig. 1).

When generating discharge instructions (di_i), we further append the brief hospital course report (bhc_i) to the clinical context. During inference, the model first generates the brief hospital course and then utilizes this output to produce the discharge instructions.

4 Experiments and Evaluation

4.1 Experimental Setup and Baselines

We fine-tune *Llama 3 8B Instruct* using the training and validation splits of the *Discharge Me!* challenge dataset. This model is chosen to maintain a fair comparison with Damm et al. (2024), who placed first on the leaderboard by employing a Dynamic Expert Selection (DES) system that included *Llama 3 8B Instruct* as one of its smaller models. For evaluation, we utilize the phase 2 test split, which was used to determine the final leaderboard rankings. The Top 3 leaderboard entries serve as a baseline for the Brief Hospital Course (BHC) and Discharge Instructions (DI) generation tasks.

We first fine-tune *Llama 3 8B Instruct* with instruction tuning (cf. Section 3.2), with prompts $prompt_i(\text{none}, \text{none})$ as a baseline for our methods, denoted as BASE. Afterwards we fine-tune *Llama 3 8B Instruct* with instruction tuning to condition the generation for user control. Since our two strategies are complementary, we experiment with different configurations. Models trained with authoring guidelines are indicated with STYLE or INSTR. respectively. Models conditioned for topic-level generation control are indicated by TOPICS.

In addition, we prompt the base model *Llama 3.1 70B Instruct* with user messages $a(c, g)$ to evaluate the overall gains of our methods.

4.2 Training Details

We fine-tune *Llama 3 8B Instruct* on 8 H100 GPUs for 3,000 steps with the AdamW 8-bit optimizer (Dettmers et al., 2022) ($\beta = (0.9, 0.999), \epsilon = 1e^{-8}$) and a batch size of 128, which amounts to training for about 2.8 epochs. We use gradient clipping with a maximum gradient norm of 1 and weight decay is set to $1e^{-4}$. Furthermore, our models are fine-tuned with instruction-tuning on completions only with rank-stabilized LoRA (Kalajdzievski, 2023) targeting all linear layers with $\alpha_{\text{LoRA}} = 64$, $\text{dropout}_{\text{LoRA}} = 0.1$, $r\text{LoRA} = 64$, and fast SVD-based PISSA (Meng et al., 2024) with 32 iterations to initialize adapter weights.

Inspired by Hu et al. (2024), who proposed to replace linear decay with a cosine cyclic decay to increase the duration of higher learning rates, we adopt a learning rate scheduler with a stable phase of 1,000 steps with learning rate $\alpha_1 = 1e^{-4}$, a decay phase of 1,800 steps corresponding to 0.25 cosine cycles to reduce the learning rate to $\alpha_2 = 5e^{-6}$, and another smoothing decay phase of 200 steps corresponding to the remaining 0.25 cosine cycles to reduce the learning rate to $\alpha_3 = 1e^{-6}$.

4.3 Evaluation

We apply nucleus sampling (top-p sampling) with a temperature of 0.1 and top-p value of 0.99 across all our models for evaluation. The structured output generated for topic-level control is stripped down to the text blocks t_i^j, which constitute the target text t_i. The performance is assessed using a comprehensive set of evaluation metrics provided by the *Discharge Me!* challenge evaluation code[3]. These metrics measure the accuracy, similarity, factual consistency, as well as the clinical relevance and correctness of the generated text with the target text.

BLEU-4 (Papineni et al., 2002) measures the precision of four-word sequences (4-grams) in the generated text against reference texts, capturing the overlap of these sequences. **ROUGE-1, ROUGE-2, ROUGE-L** (Lin, 2004) evaluate the recall of unigrams, bigrams, and the longest common subsequence between the generated and reference texts, indicating the similarity of content. **BERTScore** (Zhang et al., 2020) uses contextual

[3]https://github.com/Stanford-AIMI/discharge-me

embeddings from BERT to evaluate the semantic similarity between the generated and reference texts. **Meteor** (Banerjee and Lavie, 2005) considers synonyms and stemming to compare the generated text with reference texts, providing a more flexible measure of similarity. **AlignScore** (Zha et al., 2023) aligns generated and reference texts to measure the quality of the alignment, reflecting the coherence and consistency of the generation. **MEDCON** (wai Yim et al., 2023) is specifically designed for medical contexts, and evaluates the clinical relevance and correctness of the generated text.

4.4 Automated Evaluation of User-Controlled Text Generation

Since our methods are conditioned for user control, their evaluation requires live user input. Ideally, evaluating these models would involve a user study in which users interact with the models to generate discharge instructions and brief hospital course reports. Expert human evaluators would then rank the quality of these outputs, comparing them against the output from state-of-the-art (SOTA) systems to ensure a fair assessment. However, conducting such a study at scale is resource-intensive and time-consuming.

To address this challenge, we follow a methodology similar to previous work (Mu et al., 2024; Fakhoury et al., 2024), where LLMs are used as a proxy for human evaluators. Specifically, we utilize LLMs to simulate how the original authors of the discharge instructions and brief hospital course sections from the test set would use our models to generate their original text, the gold-standard label. This allows us to use the established evaluation framework and effectively compare our results with the *Discharge Me!* leaderboard.

For models conditioned on authoring guidelines, this process is straightforward. We pass the previously LLM-generated guidelines for the target texts along with the input to the models, assuming they are representative of human-written guidelines.

The simulation of topic-level generation control presents to be more challenging, as multi-step generation allows users to be influenced by partially written documents. In every generation step, users either accept the model-suggested topic or provide their own suggestion in form of a topic-specific heading h_i or question q_i, which is then used by the model to generate the respective text block t_i. If not satisfied, the user might redo the step. The more the user engages with the system, the higher the quality of the final output will be. For the sake of simplicity however, we will extend the user message $user_i(c, g)$ with instructions to cover the topics (or answer the questions respectively), that we extract from the topic segmentation $seg(t_i)$ of the target text, leaving more complex scenarios for future work. Since this represents a less intense engagement with the system, we expect the scores to provide a lower bound of the model performance.

5 Results and Discussion

In this section, we present the results of evaluating large language models (LLMs) on clinical text generation tasks, specifically focusing on Discharge Instructions (DIs) and Brief Hospital Courses (BHCs). We examine how incorporating topic-level generation control and authoring guidelines influences performance, comparing general instruction-tuned models to our specialized fine-tuned approach.

5.1 Performance of General Instruction-Tuned LLMs

Our evaluation of the Llama 3.1 70B Instruct model highlights the limitations of general instruction-tuned models without fine-tuning, even when provided with authoring guidelines and topic-level guidance (Table 1). These models perform significantly worse on generating Discharge Instructions (DIs) compared to Brief Hospital Courses (BHCs), despite DIs being informed by the previously generated BHCs (Table 2). This performance gap between these tasks is closed only when writing instructions are provided. However, the model performs best when provided with topic-level control, in which case providing authoring guidelines does not even further improve performance. This suggests that general models primarily benefit from content-related signals, which are less prominent and salient in authoring guidelines, rather than structural or stylistic guidance. This might be either because (a) general instruction-tuned LLMs are bad at following authoring guidelines for specialized domains, such as the clinical domain, or (b) LLMs do follow authoring guidelines, but the metrics are not sensitive to stylistic deviations. However, we explain below why the latter is not the case.

5.2 Baseline Model Performance

Our baseline model outperforms the leading models on the leaderboard across all metrics except for

	Overall	BLEU	R-1	R-2	R-L	BS	Meteor	AS	MEDCON
WisPerMed	**0.332**	0.124	0.453	0.201	0.308	0.438	0.403	0.315	0.411
HarmonAiLab@Yale	0.300	0.106	0.423	0.180	0.284	0.412	0.381	0.265	0.353
aehrc	0.297	0.097	0.414	0.192	0.284	0.383	0.398	0.274	0.332
Llama 3.1 70B Instruct									
—	0.179	0.015	0.295	0.070	0.144	0.215	0.213	0.235	0.242
STYLE	0.179	0.015	0.296	0.069	0.144	0.221	0.218	0.233	0.237
INSTR.	0.206	0.023	0.331	0.086	0.168	0.267	0.247	0.260	0.269
TOPICS	0.230	0.030	0.363	0.100	0.192	0.297	0.282	0.257	0.324
STYLE w/TOPICS	0.230	0.029	0.360	0.099	0.192	0.296	0.280	0.259	0.322
INSTR. w/TOPICS	**0.233**	0.030	0.363	0.102	0.194	0.300	0.285	0.266	0.325
BASE									
—	0.363	0.168	0.483	0.255	0.345	0.472	0.362	0.359	0.460
STYLE	0.399	0.202	0.526	0.289	0.382	0.508	0.404	0.383	0.495
INSTR.	0.420	0.224	0.547	0.310	0.404	0.527	0.428	0.403	0.515
TOPICS	0.403	0.195	0.524	0.287	0.384	0.503	0.414	0.402	0.517
STYLE w/TOPICS	0.436	0.226	0.562	0.319	0.421	0.539	0.444	0.429	0.548
INSTR. w/TOPICS	**0.445**	0.238	0.571	0.327	0.429	0.548	0.463	0.426	0.556

Table 1: Evaluation results of the *Discharge Me!* challenge leaderboard leaders WisPerMed (Damm et al., 2024), HarmonAiLab@Yale (Socrates et al., 2024) and aehrc (Liu et al., 2024), and our models on the *Discharge Me!* test set (phase 2). We refer to the Appendix, Table 3 for more comprehensive results.

	Overall	DI	BHC
Llama 3.1 70B Instruct			
—	0.179	0.153	0.204
STYLE	0.179	0.154	0.205
INSTR.	0.206	0.196	0.216
TOPICS	0.230	0.228	0.233
STYLE w/TOPICS	0.230	0.228	0.232
INSTR. w/TOPICS	0.233	0.233	0.234
BASE			
—	0.363	0.393	0.333
STYLE	0.399	0.448	0.350
INSTR.	0.420	0.476	0.364
TOPICS	0.403	0.451	0.356
STYLE w/TOPICS	0.436	0.487	0.385
INSTR. w/TOPICS	0.445	0.500	0.390

Table 2: Overall model performance for the distinct clinical text generation tasks. A more detailed break down of these scores can be found in the Appendix, Table 3.

METEOR (Table 1). Given that the leaderboard leaders have also explored fine-tuning the Llama 3 8B Instruct model using similar approaches, this result is surprising. Damm et al. (2024) fine-tune Llama 3 8B Instruct separately for each task for 3 epochs using the entire discharge summary as input. Liu et al. (2024) fine-tuned Llama 3 8B Instruct and the base model Llama 3 8B separately, but also jointly on both tasks for 5 epochs while exploring different input configurations. They reported that fine-tuning on the entire discharge summary with radiology reports diminished model performance, concluding that short and highly relevant input contexts are required to perform well. However, our model performs better, despite passing the entire discharge summary along with the radiology reports, while training for less than 3 epochs. We therefore attribute our model's success to one key factor: a more optimized training process that features longer durations of higher learning rates.

We further observe that our model performs well especially on the DI task, reaching a score of 0.393 (Table 2). This might either be due to less stylistic variations in DIs as opposed to BHCs, or due to the DI generation step being informed by the generated BHC, suggesting that such pre-summarization steps can significantly enhance overall task performance in LLMs, particularly when fine-tuned for such sequences.

5.3 Impact of Authoring Guidelines

Providing authoring guidelines further enhances the performance of our models (Table 1). Notably, style guidelines already compensate for the deficit in METEOR scores, and even lead to improvements in metrics related to factual consistency and clinical correctness, such as AlignScore and MEDCON. Writing instructions, which provide more explicit guidance than style guidelines alone, further enhance the model's performance. This indi-

cates that generating instructive authoring guidelines with LLMs result in more salient guidance signals. Overall, this approach shows that authoring guidelines do effectively help aligning text generation with user expectations. However, further investigation is needed to determine the specific types of authoring guidelines that are most beneficial and how they complement content selection.

5.4 Influence of Topic-Level Generation Control

Conditioning for topic-level generation control improves our baseline model's performance as well, matching the results achieved with style guides as shown in Table 1. When authoring guidelines are added on top of topic-level control, performance increases further across all metrics. Our best model **BASE w/instr w/hq** performs particularly well on the DI generation task, with some metrics, such as ROUGE-1 (0.612), BERTScore (0.587), and MEDCON (0.594), reaching particularly high scores in DI generation (Table 3).

Contrasting this with the evaluation of general instruction-tuned models, we can conclude that (a) our authoring guidelines contain salient guidance signals beyond content selection, most likely concerning the writing style or the structural arrangement of the topics, and (b) the necessity of conditioning LLMs for such user control, since general instruction-tuned LLMs were not able to pick up on these guidance signals. Furthermore, the results show the sensitivity of all metrics to stylistic deviations and underscore the necessity of complementing traditional datasets with stylistic guidance for task performance and extending the task definition for clinical text generation.

6 Conclusion and Future Work

In this work, we proposed strategies for conditioning clinical text generation using LLMs to provide clinicians with more control over both content and style. Our approach addresses critical challenges in healthcare AI by offering greater transparency and user involvement in the generation process, bridging the gap between automation and the need for human oversight in clinical workflows. By incorporating topic-level structured generation and authoring guidelines, we demonstrated significant improvements in the accuracy, clinical relevance, and factual consistency of generated clinical texts. However, our current evaluation reflects a lower bound on model performance, as we did not simulate live user interaction to dynamically guide the generation process. Future research could explore more sophisticated approaches for providing real-time guidance signals, informed by ongoing generation progress, simulating how users refine outputs in a more interactive manner.

Topic-level generation control introduces a modularity that offers opportunities for further enhancement. For example, by selectively filtering input data and integrating domain-specific knowledge on a topic-by-topic basis, each generation step could be informed solely by the relevant information required. Moreover, future work could explore alternative ways of structuring the output to further improve its flexibility and accuracy.

While our augmented dataset with topic segmentation and authoring guidelines has shown promising results, further research is needed to refine these techniques. Specifically, two key questions remain: (a) what is the optimal granularity for multi-step generation control? and (b) which signals in the authoring guidelines most significantly impact the quality and relevance of clinical text generation?

Despite our automated evaluation methods, user studies remain essential to fully assess the practical impact of AI-assisted, topic-specific generation on clinician workloads, to determine whether this approach effectively reduces workloads without adding new cognitive demands.

Finally, there is considerable potential for expanding the system's personalization capabilities. Future research could investigate alternative augmentation strategies, such as automatically detecting stylistic preferences based on clinician profiles, context-sensitive generation tailored to specific medical departments or practices, or automated extraction of authoring guideline templates for different clinical text generation use cases. These enhancements would make the system easier to adapt and versatile, ensuring it meets the unique stylistic and content needs of different clinical environments without increasing the workload. Nonetheless, it is important to remember that the fundamental issues—such as hallucinations and factual errors—are not resolved simply by controlled generation. Users may overestimate the system's capabilities, as the output may appear to meet their expectations, potentially leading to inattentional blindness toward errors. Further research is needed to prevent risks of unintentional blindness and to develop methods that ensure more factually consistent outputs.

7 Limitations

Firstly, we evaluated only a single instruction-tuned model, *Llama 3.1 70B Instruct*. Larger and more powerful models, such as *Llama 3 405B Instruct* and *GPT-4*, may exhibit different performance on the tasks and augmentations explored in this paper. These larger models often demonstrate new capabilities that emerge with increased scale (Zoph et al., 2022), depending on the evaluation metrics used (Schaeffer et al., 2023).

Additionally, we used the same LLM for dataset augmentation to generate authoring guidelines and perform topic segmentation. While we carefully engineered the prompts to reduce answer leakage and improve output quality, we have not fully evaluated the quality of these outputs to determine if LLMs can adequately replace human annotators for such tasks. However, this limitation may become less significant with the future use of more powerful LLMs, which could automatically improve the quality of these annotations.

Furthermore, the performance of our models is closely tied to the quality and structure of the training data, which may not fully capture the diversity of real-world clinical documentation styles and practices. Although we enhanced the dataset with topic segmentation and authoring guidelines, these methods may not generalize across all clinical domains or document types. Different specialties or clinical settings may require alternative approaches, and expanding the dataset to include a broader range of writing styles and conventions will be crucial for ensuring wider applicability.

Our evaluation primarily relied on automated metrics and simulated user interactions. While these methods provide valuable initial insights, they do not fully capture the complexity of human preferences, real-time decision-making, or the specific needs of clinicians in practice. To fully validate the system's effectiveness and usability, comprehensive user studies involving live clinician interactions will be necessary.

References

Omar Alfarghaly, Rana Khaled, Abeer Elkorany, Maha Helal, and Aly Fahmy. 2021. Automated radiology report generation using conditioned transformers. *Informatics in Medicine Unlocked*, 24:100557.

Kenichiro Ando, Takashi Okumura, Mamoru Komachi, Hiromasa Horiguchi, and Yuji Matsumoto. 2022. Is artificial intelligence capable of generating hospital discharge summaries from inpatient records? *PLOS Digital Health*, 1(12):e0000158.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Reece Alexander James Clough, William Anthony Sparkes, Oliver Thomas Clough, Joshua Thomas Sykes, Alexander Thomas Steventon, and Kate King. 2024. Transforming healthcare documentation: harnessing the potential of ai to generate discharge summaries. *BJGP open*, 8(1).

Amin Dada, Marie Bauer, Amanda Butler Contreras, Osman Alperen Koraş, Constantin Marc Seibold, Kaleb E Smith, and Jens Kleesiek. 2024. Does biomedical training lead to better medical performance? *arXiv preprint arXiv:2404.04067*.

Hendrik Damm, Tabea Margareta Grace Pakull, Bahadır Eryılmaz, Helmut Becker, Ahmad Idrissi-Yaghir, Henning Schäfer, Sergej Schultenkämper, and Christoph M. Friedrich. 2024. WisPerMed at "discharge me!": Advancing text generation in healthcare with large language models, dynamic expert selection, and priming techniques on MIMIC-IV. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 105–121, Bangkok, Thailand. Association for Computational Linguistics.

Tim Dettmers, Mike Lewis, Sam Shleifer, and Luke Zettlemoyer. 2022. 8-bit optimizers via block-wise quantization. *Preprint*, arXiv:2110.02861.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, and Ahmad Al-Dahle et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Daniel Dubinski, Sae-Yeon Won, Svorad Trnovec, Bedjan Behmanesh, Peter Baumgarten, Nazife Dinc, Juergen Konczalla, Alvin Chan, Joshua D Bernstock, Thomas M Freiman, et al. 2024. Leveraging artificial intelligence in neurosurgery—unveiling chatgpt for neurosurgical discharge summaries and operative reports. *Acta neurochirurgica*, 166(1):38.

Samuel T Edwards, Pamela M Neri, Lynn A Volk, Gordon D Schiff, and David W Bates. 2014. Association of note quality and quality of care: a cross-sectional study. *BMJ quality & safety*, 23(5):406–413.

Simon Ellershaw, Christopher Tomlinson, Oliver E Burton, Thomas Frost, John Gerrard Hanrahan, Danyal Zaman Khan, Hugo Layard Horsfall, Mollie Little, Evaleen Malgapo, Joachim Starup-Hansen, et al. 2024. Automated generation of hospital discharge summaries using clinical guidelines and large language models. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*.

Sarah Fakhoury, Aaditya Naik, Georgios Sakkas, Saikat Chakraborty, and Shuvendu K. Lahiri. 2024. Llm-based test-driven interactive code generation: User study and empirical evaluation. *IEEE Transactions on Software Engineering*, 50(9):2254–2268.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Vince Hartman and Thomas R Campion. 2022. A day-to-day approach for automating the hospital course section of the discharge summary. *AMIA Summits on Translational Science Proceedings*, 2022:216.

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.

Takanobu Hirosawa, Yukinori Harada, Masashi Yokose, Tetsu Sakamoto, Ren Kawamura, and Taro Shimizu. 2023. Diagnostic accuracy of differential-diagnosis lists generated by generative pretrained transformer 3 chatbot for clinical vignettes with common chief complaints: A pilot study. *International Journal of Environmental Research and Public Health*, 20(4).

Leora I. Horwitz, John P. Moriarty, Christine Chen, Robert L. Fogerty, Ursula C. Brewster, Sandhya Kanade, Boback Ziaeian, Grace Y. Jenq, and Harlan M. Krumholz. 2013. Quality of Discharge Practices and Patient Understanding at an Academic Medical Center. *JAMA Internal Medicine*, 173(18):1715–1722.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, Ning Ding, Chao Jia, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *Preprint*, arXiv:2404.06395.

Thomas Huang, Conrad W Safranek, Vimig Socrates, David Chartash, Donald Wright, Monisha Dilip, Rohit B. Sangal, and Richard Andrew Taylor. 2024. Patient-representing population's perceptions of gpt-generated versus standard emergency department discharge instructions: Randomized blind survey assessment. *Journal of Medical Internet Research*, 26.

Gretchen M Hultman, Jenna L Marquard, Elizabeth Lindemann, Elliot Arsoniadis, Serguei Pakhomov, and Genevieve B Melton. 2019. Challenges and opportunities to improve the clinician experience reviewing electronic progress notes. *Applied clinical informatics*, 10(03):446–453.

Fantine Huot, Joshua Maynez, Shashi Narayan, Reinald Kim Amplayo, Kuzman Ganchev, Annie Priyadarshini Louis, Anders Sandholm, Dipanjan Das, and Mirella Lapata. 2023. Text-blueprint: An interactive platform for plan-based conditional generation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 105–116, Dubrovnik, Croatia. Association for Computational Linguistics.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Damjan Kalajdzievski. 2023. A rank stabilization scaling factor for fine-tuning with lora. *Preprint*, arXiv:2312.03732.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Zichao Lin, Shuyan Guan, Wending Zhang, Huiyan Zhang, Yugang Li, and Huaping Zhang. 2024. Towards trustworthy llms: a review on debiasing and dehallucinating in large language models. *Artif. Intell. Rev.*, 57:243.

Jialin Liu, Changyu Wang, and Siru Liu. 2023. Utility of chatgpt in clinical practice. *Journal of Medical Internet Research*, 25:e48568.

Jinghui Liu, Aaron Nicolson, Jason Dowling, Bevan Koopman, and Anthony Nguyen. 2024. e-health CSIRO at "discharge me!" 2024: Generating discharge summary sections with fine-tuned language

models. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 675–684, Bangkok, Thailand. Association for Computational Linguistics.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. 2024. Pissa: Principal singular values and singular vectors adaptation of large language models. *Preprint*, arXiv:2404.02948.

Bertalan Meskó and Eric J. Topol. 2023. The imperative for regulatory oversight of large language models (or generative ai) in healthcare. *NPJ Digital Medicine*, 6.

Fangwen Mu, Lin Shi, Song Wang, Zhuohao Yu, Binquan Zhang, ChenXue Wang, Shichao Liu, and Qing Wang. 2024. Clarifygpt: A framework for enhancing llm-based code generation via requirements clarification. *Proc. ACM Softw. Eng.*, 1(FSE).

Shashi Narayan, Joshua Maynez, Reinald Kim Amplayo, Kuzman Ganchev, Annie Louis, Fantine Huot, Anders Sandholm, Dipanjan Das, and Mirella Lapata. 2023. Conditional generation with a question-answering blueprint. *Transactions of the Association for Computational Linguistics*, 11:974–996.

Jesutofunmi A Omiye, Haiwen Gui, Shawheen J Rezaei, James Zou, and Roxana Daneshjou. 2024. Large language models in medicine: The potentials and pitfalls : A narrative review. *Annals of internal medicine*, 177(2):210—220.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, and Ilge Akkaya et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Sajan B Patel and Kyle Lam. 2023. Chatgpt: the future of discharge summaries? *The Lancet Digital Health*, 5(3):e107–e108.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3419–3448, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Craige Roberts. 2012. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5(6):1–69.

Lorenzo Ruinelli, Amos Colombo, Mathilde Rochat, Sotirios Georgios Popeskou, Andrea Franchini, Sandra Mitrović, Oscar William Lithgow, Joseph Cornelius, and Fabio Rinaldi. 2024. Experiments in automated generation of discharge summaries in Italian. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024*, pages 137–144, Torino, Italia. ELRA and ICCL.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems*, volume 36, pages 55565–55581. Curran Associates, Inc.

Thomas Searle, Zina Ibrahim, James Teo, and Richard JB Dobson. 2023. 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.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, Mike Schaekermann, Amy Wang, Mohamed Amin, Sami Lachgar, Philip Mansfield, Sushant Prakash, Bradley Green, Ewa Dominowska, Blaise Aguera y Arcas, Nenad Tomasev, Yun Liu, Renee Wong, Christopher Semturs, S. Sara Mahdavi, Joelle Barral, Dale Webster, Greg S. Corrado, Yossi Matias, Shekoofeh Azizi, Alan Karthikesalingam, and Vivek Natarajan. 2023. Towards expert-level medical question answering with large language models. *Preprint*, arXiv:2305.09617.

Vimig Socrates, Thomas Huang, Xuguang Ai, Soraya Fereydooni, Qingyu Chen, R Andrew Taylor, and David Chartash. 2024. Yale at "discharge me!": Evaluating constrained generation of discharge summaries with unstructured and structured information. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 724–730, Bangkok, Thailand. Association for Computational Linguistics.

Mohsen Soleimani, Navisa Seyyedi, Seyed Mohammad Ayyoubzadeh, Sharareh Rostam Niakan Kalhori, and Hamidreza Keshavarz. 2024. Practical evaluation of chatgpt performance for radiology report generation. *Academic Radiology*.

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, et al. 2023. Evaluating large language models on medical evidence summarization. *NPJ digital medicine*, 6(1):158.

Zhen Tao, Dinghao Xi, Zhiyu Li, Liumin Tang, and Wei Xu. 2024. Cat-llm: Prompting large language models with text style definition for chinese article-style transfer. *arXiv preprint arXiv:2401.05707*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, and Radu Soricut et al. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, and Amjad Almahairi et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Jan Van Kuppevelt. 1995. Discourse structure, topicality and questioning. *Journal of linguistics*, 31(1):109–147.

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. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142.

Wen wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Scientific Data*, 10.

Zhanyu Wang, Lingqiao Liu, Lei Wang, and Luping Zhou. 2023. R2gengpt: Radiology report generation with frozen llms. *Meta-Radiology*, 1(3):100033.

Martin Chieng Were, Changyu Shen, Mwebesa B. Bwana, Nneka Emenyonu, Nicholas Musinguzi, Frank Nkuyahaga, Annet Kembabazi, and William M. Tierney. 2010. Creation and evaluation of emr-based paper clinical summaries to support hiv-care in uganda, africa. *International journal of medical informatics*, 79 2:90–6.

Justin Xu, Zhihong Chen, Andrew Johnston, Louis Blankemeier, Maya Varma, Jason Hom, William J. Collins, Ankit Modi, Robert Lloyd, Benjamin Hopkins, Curtis Langlotz, and Jean-Benoit Delbrouck. 2024. Overview of the first shared task on clinical text generation: RRG24 and "discharge me!". In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 85–98, Bangkok, Thailand. Association for Computational Linguistics.

Kevin Yang and Dan Klein. 2021. FUDGE: Controlled text generation with future discriminators. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3511–3535, Online. Association for Computational Linguistics.

Shuxin Yang, Xian Wu, Shen Ge, Zhuozhao Zheng, S Kevin Zhou, and Li Xiao. 2023. Radiology report generation with a learned knowledge base and multi-modal alignment. *Medical Image Analysis*, 86:102798.

Jonah Zaretsky, Jeong Min Kim, Samuel Baskharoun, Yunan Zhao, Jonathan S 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.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. AlignScore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348, Toronto, Canada. Association for Computational Linguistics.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2023. A survey of controllable text generation using transformer-based pre-trained language models. *ACM Computing Surveys*, 56(3):1–37.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *Preprint*, arXiv:1904.09675.

Barret Zoph, Colin Raffel, Dale Schuurmans, Dani Yogatama, Denny Zhou, Don Metzler, Ed H. Chi, Jason Wei, Jeff Dean, Liam B. Fedus, Maarten Paul Bosma, Oriol Vinyals, Percy Liang, Sebastian Borgeaud, Tatsunori B. Hashimoto, and Yi Tay. 2022. Emergent abilities of large language models. *TMLR*.

Xu Zou, Da Yin, Qingyang Zhong, Hongxia Yang, Zhilin Yang, and Jie Tang. 2021. Controllable generation from pre-trained language models via inverse prompting. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2450–2460.

A Appendix

A.1 Supplementary Results

	Task	Overall	BLEU	R-1	R-2	R-L	BS	Meteor	AS	MEDCON
Llama 3.1 70B Instruct										
—	BHC	0.204	0.020	0.333	0.092	0.175	0.278	0.203	0.229	0.303
	DI	0.153	0.009	0.257	0.048	0.113	0.153	0.223	0.242	0.182
STYLE	BHC	0.205	0.021	0.333	0.089	0.173	0.288	0.209	0.231	0.292
	DI	0.154	0.009	0.259	0.050	0.115	0.153	0.226	0.234	0.183
INSTR.	BHC	0.216	0.028	0.357	0.101	0.188	0.303	0.226	0.220	0.308
	DI	0.196	0.017	0.305	0.072	0.147	0.231	0.268	0.299	0.230
TOPICS	BHC	0.233	0.033	0.372	0.106	0.196	0.307	0.267	0.235	0.348
	DI	0.228	0.027	0.354	0.094	0.189	0.286	0.296	0.278	0.300
STYLE w/TOPICS	BHC	0.232	0.032	0.370	0.105	0.196	0.307	0.264	0.237	0.346
	DI	0.228	0.027	0.351	0.093	0.188	0.284	0.297	0.281	0.299
INSTR. w/TOPICS	BHC	0.234	0.033	0.372	0.106	0.198	0.309	0.266	0.238	0.349
	DI	0.233	0.028	0.355	0.097	0.191	0.291	0.304	0.294	0.302
BASE										
—	BHC	0.333	0.142	0.465	0.228	0.313	0.460	0.335	0.290	0.435
	DI	0.393	0.193	0.502	0.283	0.377	0.484	0.390	0.428	0.484
STYLE	BHC	0.350	0.158	0.488	0.242	0.330	0.477	0.356	0.297	0.452
	DI	0.448	0.247	0.565	0.337	0.434	0.539	0.452	0.469	0.538
INSTR.	BHC	0.364	0.171	0.505	0.255	0.343	0.489	0.376	0.304	0.470
	DI	0.476	0.277	0.589	0.366	0.464	0.565	0.480	0.503	0.560
TOPICS	BHC	0.356	0.149	0.487	0.239	0.333	0.471	0.369	0.317	0.482
	DI	0.451	0.240	0.561	0.336	0.436	0.535	0.459	0.487	0.552
STYLE w/TOPICS	BHC	0.385	0.178	0.524	0.268	0.367	0.504	0.396	0.332	0.513
	DI	0.487	0.273	0.600	0.370	0.474	0.574	0.491	0.526	0.584
INSTR. w/TOPICS	BHC	0.390	0.183	0.530	0.271	0.370	0.509	0.410	0.327	0.517
	DI	0.500	0.292	0.612	0.383	0.488	0.587	0.517	0.525	0.594

Table 3: The average scores per metrics for our evaluations, broken down by the two tasks: discharge instructions (DI) generation and brief hospital course (BHC) generation.