Towards Conditioning Clinical Text Generation for User Control

Osman Alperen Koraş[1] **Rabi Bahnan**[1] **Jens Kleesiek**[1,2,3,4] **Amin Dada**[1]
[1]Institute for AI in Medicine (IKIM), University Hospital Essen (AöR), Essen, Germany
[2]Cancer Research Center Cologne Essen (CCCE), West German Cancer Center Essen
University Hospital Essen (AöR), Essen, Germany
[3]German Cancer Consortium (DKTK, Partner site Essen), Heidelberg, Germany
[4]Department of Physics, TU Dortmund, Dortmund, Germany

Abstract

Deploying natural language generation systems in clinical settings remains challenging despite advances in Large Language Models (LLMs), which continue to exhibit hallucinations and factual inconsistencies, necessitating human oversight. This paper explores automated dataset augmentation using LLMs as human proxies to condition LLMs for clinician control without increasing cognitive workload. On the BioNLP ACL'24 Discharge Me! Shared Task, we achieve new state-of-the-art results with simpler methods than prior submissions through more efficient training, yielding a 9% relative improvement without augmented training and up to 34% with dataset augmentation. Preliminary human evaluation further supports the effectiveness of our approach, highlighting the potential of augmenting clinical text generation for control to enhance relevance, accuracy, and factual consistency.

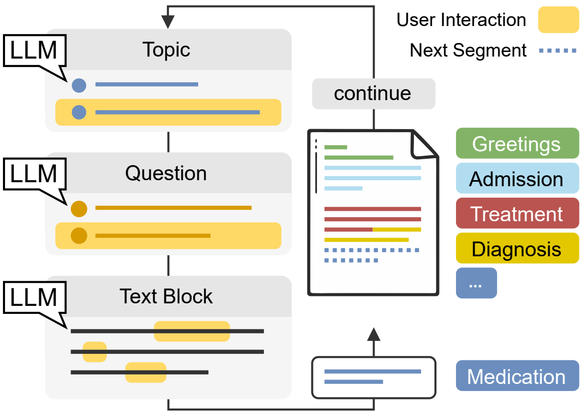

Figure 1: An interactive workflow showcasing topic-level generation control. The LLM is prompted once with the respective context to begin structured generation. After each element, generation is paused, enabling users to sequentially refine content by editing LLM-suggested topic headings, questions, and text blocks. The generation resumes with user-verified content.

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, LLMs offer potential for automating 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, deploying AI in clinical settings remains a critical challenge due to the high cost of hallucinations, factual inconsistencies, and misinterpretations (Ji et al., 2023; Lin et al., 2024; Tang et al., 2023; Dada et al., 2024). Even minor inaccuracies in AI-generated clinical content can lead to severe consequences, such as misdiagnoses, incorrect treatments, or harmful patient outcomes. Ethical considerations further complicate this process, calling for clinicians to hold accountability for medical decisions through rigorous oversight (Meskó and Topol, 2023; Omiye et al., 2024). At the same time, verifying AI-generated content introduces new cognitive burdens, potentially negating the intended efficiency gains of automation. As clinicians already face high cognitive workloads, addressing this paradox is essential to harness AI's potential in clinical settings without increasing risks or workloads. To strike this balance, AI systems must provide clinicians with control and transparency, ensuring outputs align with clinical contexts, communication styles, and guidelines. This paper explores whether augmenting traditional datasets to condition LLMs for controlled clinical text generation is a viable solution. Specifically, we introduce a system that

Findings of the Association for Computational Linguistics: ACL 2025, pages 10549–10569
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

separates stylistic and content-related requirements, breaking down generation into distinct, manageable writing subtasks. This reduces the complexity of content creation and human verification through a separation of concerns, empowering users to impose authoring guidelines and dynamically guide the process while moving away from black-box models that limit clinician involvement.

Since traditional datasets do not inherently support such user control, we augment them with authoring guidelines and topic segmentation to condition models for style and content control. Automated evaluation suggests that our approach significantly enhances relevance, accuracy, and factual consistency, highlighting the potential of such augmentations for clinical text generation. Furthermore, we find that traditional instruction-tuning for clinical text generation can be significantly improved through optimized hyperparameter settings, without increasing the compute budget. Our key contributions are:

New state-of-the-art. We set a new state-of-the-art on the BioNLP ACL'24 Shared Task 'Discharge Me!' challenge through efficient training, while being simpler and requiring less training compute.

Dataset Augmentation. We propose methods[1] using LLMs as human proxies to augment traditional datasets, enabling granular control over content and style in clinical text generation. This yields a 34% relative improvement over prior state-of-the-art, representing a lower bound on potential gains.

Human Evaluation. We conduct preliminary human evaluation, validating the effectiveness of our approach. Our findings suggest that LLMs show great potential as proxies for human annotators in complex clinical tasks.

2 Related Work

In recent years, research on LLM-based clinical text generation has grown, demonstrating promise in generating discharge summaries (Ando et al., 2022; Ellershaw et al., 2024; Clough et al., 2024; Dubinski et al., 2024), brief hospital courses (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). Notably, some studies report that physicians often prefer AI-generated clinical texts over manually written ones (Van Veen et al., 2024). Despite this progress, most approaches treat clinical text generation as an end-to-end task, limiting user control and intervention. A recent example is the BioNLP ACL'24 Shared Task 'Discharge Me!' (Xu et al., 2024), targeting discharge summary generation. However, the complexity of clinical texts, which often require external sources of information and are subject to individual guidelines and writing styles, makes end-to-end generation less feasible in practice, and points to the need for more flexible generation allowing users to control specific aspects of the output, such as content and style.

Controlled text generation (CTG) addresses this by allowing users to steer outputs through specific conditions—such as tone, structure, or terminology—while maintaining fluency and relevance (Zhang et al., 2023). Prior work in this area has explored different control mechanisms, 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 by framing clinical documents as responses to implicit questions arising from their intended purpose. These questions are typically addressed in a structured manner, even when the document appears unstructured. Using fine-grained topic segmentation, we aim to uncover this underlying structure by generating headings and QUDs, with corresponding text segments acting as their answers. This approach aligns topics with specific writing subtasks, simplifying the generation process while preserving the inherent structure of clinical documentation.

3 Conditioning Clinical Text Generation for User Control

We explore two strategies to condition Large Language Models (LLMs) for controlled clinical text generation: (a) topic-level structured generation and (b) authoring guidelines. However, implement-

[1] https://github.com/TIO-IKIM/controlled-clinical-generation

ing these strategies reveal limitations in traditional datasets, particularly in clinical text generation.

3.1 Limitations in Traditional Datasets

Traditionally, training datasets are built on the assumption that more data leads to better generalization. However, in conditional text generation (e.g., summarization) the same task can be completed in multiple stylistically distinct but equally valid ways. Despite this, evaluation benchmarks typically provide only a single reference text, failing to account for the diversity of valid solutions or specifying which variant of task completion is expected. This issue is made by design and cannot be resolved simply by increasing dataset size. Consequently, models are evaluated against a single stylistic realization of a task, potentially skewing evaluation results.

This is particularly evident in clinical datasets, where medical documents exhibit significant differences in quality, format, and style — even within the same task (Pollard et al., 2013; Edwards et al., 2014; Hultman et al., 2019). Discharge summaries, for instance, are often compiled from pre-existing records authored by multiple individuals. Contents are often copied across teams, departments, or wards, each adhering to distinct conventions shaped by institutional workflows, time constraints, and resource limitations, leading to inherent stylistic inconsistencies, which is further amplified by situational pressures. Moreover, medical professionals exhibit highly distinctive writing styles, often to the extent that colleagues can recognize one another solely by their writings. Consequently, even within a single discharge summary, different sections may reflect different writing styles, making it impossible to reliably infer the appropriate writing style for one section from the remaining document.

This issue has been largely overlooked in prior research, and to our knowledge, no systematic study has investigated its implications. In particular, the extent to which evaluation metrics are sensitive to stylistic variations, and the degree to which stylistic features emerge due to spurious correlations in input data, remains unclear. To ensure models can be held accountable for stylistic deviations, we extend the task definition by integrating authoring guidelines into the input context, conditioning the model to adhere to explicit stylistic requirements. This introduces a clear separation of concerns: synthesizing clinically relevant information to complete the task (**content**) and ensuring conformity to specified conventions (**style**). Moreover, explicitly conditioning models on authoring guidelines facilitates the emergence of stylistic features through user control, rather than spurious correlations, enabling clinicians to specify institutional or personalized guidelines during inference and promising better generalization.

Another limitation with traditional datasets is their end-to-end design, where the entire output is generated in a single step from the input. This inherently restricts user intervention and control during generation. To train models for (a) controllable and (b) intervenable generation (cf. Fig. 1), we need models to sequentially generate output block by block in a structured format with (a) guidance signals to steer the generation of individual blocks and (b) control sequences to start and terminate individual blocks. To address this, we explore fine-grained topic segmentation to structure target texts t_i into XML-formatted sequences.

3.2 Topic-Level Generation Control

To train models for controllable and intervenable generation, we tasked Llama 3.1 70B Instruct with fine-grained topic segmentation of target texts t_i. The LLM is prompted to segment texts $t_i = (t_i^1, ..., t_i^n)$ into smaller text blocks t_i^k, while generating topic-specific headings $\mathring{h}_i^k$ and questions $\mathring{q}_i^k$ for each segment. The output is requested as an XML-structured sequence

$$\mathring{seg}(t_i) = \left[\mathring{h}_i^1, \mathring{q}_i^1, \mathring{t}_i^1, ..., \mathring{h}_i^n, \mathring{q}_i^n, \mathring{t}_i^n\right],$$

in the following format:

```
<topic>$\mathring{h}_i^1$</topic>
<question>$\mathring{q}_i^1$</question>
<span>$\mathring{t}_i^1$</span>
...
<topic>$\mathring{h}_i^n$</topic>
<question>$\mathring{q}_i^n$</question>
<span>$\mathring{t}_i^n$</span>
```

While the headings and questions serve as guidance signals during generation, the XML tags serve as control sequences to stop, adjust and continue generation in each distinct phase (Fig. 1). The prompt (Tab. 10) is designed to enforce fine-grained topic segmentation, without imposing a particular concept of topics or questions. It's summarized as follows: (1) a new segment should begin when the clinical focus changes, which we associate with a new writing subtask, (2) headings $\mathring{h}_i^k$

should summarize their respective segment, which we equate with the topic, and which (3) should be rephrased as a question $\mathring{q}_i^k$ answered by the respective segment t_i^k, which we consider to be the Question Under Discussion (QUD) of said segment. The remaining guidelines are provided to ensure standardization. It is important to note that the number of topics and questions generated is not a predefined hyperparameter but rather emerges implicitly from the LLM-driven segmentation process, reflecting the inherent structure and information density of the input content.

A post-processing step then restores the original character sequences $t_i^1, ..., t_i^n$ from t_i for each generated text block $\mathring{t}_i^1, \ldots, \mathring{t}_i^n$ (see Appendix D), as the LLM generated text blocks $\mathring{t}_i^k$ may not replicate the input t_i. The final output is denoted as:

$$seg(t_i) = \left[\mathring{h}_i^1, \mathring{q}_i^1, t_i^1, ..., \mathring{h}_i^n, \mathring{q}_i^n, t_i^n \right].$$

However, to avoid introducing inconsistencies between headings, questions, and text blocks, this step is applied selectively only to those segmentations $\mathring{seg}(t_i)$, which introduce only minor alterations to the input (see Appendix D).

3.3 Authoring Guidelines

From a user perspective, authoring guidelines govern the requirements a document must comply with. These may range from stylistic features to structural constraints. Conditioning text generation on such guidelines may therefore not only improve alignment of model outputs with user intent, but also provide greater control over generation. However, traditional datasets often lack such guidelines. In this work, we explore the feasibility of using LLMs to close this gap in clinical datasets.

Specifically, we explore the use of two types of automatically generated authoring guidelines for clinical documents t_i, which differ in their formulation: (a) style guidelines, which describe the stylistic features a clinical document should express and (b) writing instructions, guiding a non-specialist in writing a clinical document that serves the intended purpose while expressing the desired stylistic features. To achieve this, Llama 3.1 70B Instruct is prompted independently for each target text t_i as follows:

Style Guidelines. The LLM is prompted to describe the stylistic features of the target text t_i, including tone, document format, layout, composition, text structure, use of language (including abbreviations and medical jargon), and intended audience (cf. Tab. 11).

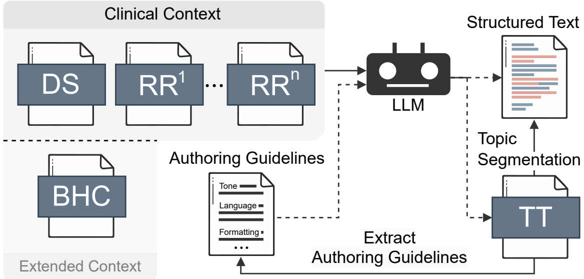

Figure 2: Instruction-tuning pipeline. Dashed lines indicate paths that depend on the training configuration. Models with topic-level control are trained to generate XML-structured text. The extended context is provided only for TT = DI. Abbreviations: Discharge Summary (DS), Radiology Report (RR), Discharge Instructions (DI), Brief Hospital Course (BHC), Target Text (TT).

Writing Instructions. The LLM is prompted to generate markdown-formatted instructions for guiding a non-specialist in replicating the target text t_i, including directives on the same stylistic features as above while specifying the purpose, document type and outline (cf. Tab. 12).

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.

3.4 Instruction Tuning for Controlled Clinical Text Generation

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. We selected this challenge as a principled and practical benchmark: it is built on the MIMIC-IV dataset (?), which, to our knowledge, remains the only large-scale, publicly available clinical dataset suitable for structured summarization and conditioned text generation in real clinical contexts. The challenge is grounded in two distinct and clinically meaningful generation tasks and provides a validated suite of evaluation metrics (Xu et al., 2024). Furthermore, working with this benchmark enables us to open-source our code, data, and methods, fostering reproducibility and allowing practitioners to adapt our approach to their own institutional data, which we consider essential

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

User Message

```
{{discharge summary}}
{{radiology report 1}}
          ...
{{radiology report n}}
{{brief hospital course}}
{{authoring guidelines}}
{{instructions}}
```

Assistant Message

```
{{output}}
```

Figure 3: The generic template for $prompt_i(c, g)$ used for instruction-tuning.

for responsible research in sensitive domains like healthcare. 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.

Dataset. The dataset consists of 109,168 discharge summaries from the MIMIC-IV dataset, each containing a Brief Hospital Course (BHC) and a Discharge Instructions (DI) section. It is divided into training (68,785), validation (14,719), phase I test (14,702), and phase II test (10,962) sets. 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. The DI and BHC sections are removed from the discharge summary, and serve as target texts t_i. The clinical input constitutes of the remaining discharge summary (DS) and radiology reports (RR).

To address the aforementioned limitations (3.1), we generate topic segmentations (3.2), style guidelines and writing instructions (3.3) for each DI and BHC section t_i separately. We employ Llama 3.1 70B Instruct for these tasks, as LLMs have shown to be an effective substitute for human annotators (Gilardi et al., 2023; Perez et al., 2022).

Instruction-Tuning Prompts. We fine-tune our models with instruction-tuning on completions only using a generic template (cf. Fig. 3)

$$prompt_i(c, g) = (user_i(c, g), assistant_i(c)),$$

where $c \in \{\texttt{none, topics}\}$ denotes the possible configurations for structuring the generation output for control and $g \in \{\texttt{none, style, instr}\}$ denotes the possible configurations for using authoring guidelines.

User Messages. $user_i(c, g)$ include the clinical context, consisting of the discharge summary ds_i and radiology reports $r_i^1, \ldots, r_i^j$. For generating discharge instructions (di_i), we additionally include the brief hospital course report (bhc_i). If $g \in \{\texttt{style, instr}\}$, we also include the respective authoring guidelines (cf. Fig. 2) and instruct the model to comply. If $c = \texttt{topics}$, the model is instructed to generate XML-structured output for topic-level structured generation. Separate instructions are provided for the DI and BHC generation tasks.

Assistant Messages. $assistant_i(c)$ contains the desired output, which is the plain target text $t_i \in \{di_i, bhc_i\}$ for $c = \texttt{none}$, or the XML-structured output $seg_i(t_i)$ for $c = \texttt{topics}$ (Sec. 3.2).

4 Experiments and Evaluation

4.1 Experimental Setup and Baselines

We fine-tune *Llama 3 8B Instruct* on the training split of the *Discharge Me!* challenge dataset with instruction tuning using $prompt_i(c, g)$ for all possible configurations (see Section 3.4). See Appendix A for training details. 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. We evaluate on the test-phase-2 split used to determine the final leaderboard rankings. The Top 3 leaderboard entries serve as the state-of-the-art baseline for the Brief Hospital Course (BHC) and Discharge Instructions (DI) generation tasks.

BASE denotes our model which is trained without any data augmentations ($c = \texttt{none}$, $g = \texttt{none}$). It serves as a baseline for our other models. Models trained with authoring guidelines ($g \in \{\texttt{style, instr}\}$) are indicated with **W/STYLE** or **W/INSTR** respectively. Similarly, models trained on structured output ($c = \texttt{topics}$) are indicated with **W/TOPICS**.

In addition, we prompt the stronger base model *Llama 3.3 70B Instruct* with user messages $user_i(c, g)$ zero-shot and three-shot to assess the gains provided by dataset augmentations without any fine-tuning.

4.2 Automated Evaluation

All our models are evaluated using the code provided by the *Discharge Me!* challenge[3], which employs a comprehensive set of metrics (see Appendix B) to assess lexical similarity (**BLEU-4**, **ROUGE-1**, **ROUGE-2**, **ROUGE-L**, **METEOR**), semantic similarity (**BERTScore**), factual consistency (**AlignScore**), and the clinical relevance and correctness (**MEDCON**) of the generated texts $\hat{t}_i$ in comparison to the gold-standard target texts t_i. AlignScore, in particular, is designed to measure factual consistency and has been shown to achieve high accuracy on benchmarks that include hallucination as a distinct error category (Zha et al., 2023), making it a reliable indicator of hallucination reduction. It was reported, that this ensemble resulted in rankings that aligned well with clinician evaluation (Xu et al., 2024). For evaluation, we first complete the BHC task and then use the output to generate the DI section. Greedy decoding is used for inference. For models W/TOPICS, which generate XML-structured outputs $seg(\hat{t}_i)$, the output is parsed into plain text by joining the spans $\hat{t}_i^1, \ldots, \hat{t}_i^n$ with white spaces to retrieve the final model output.

To simulate user-control, we adopt a methodology (see Appendix B) inspired by prior work (Mu et al., 2024; Fakhoury et al., 2024), leveraging LLMs as proxies for human evaluators to automate evaluation on existing benchmarks. Specifically, an LLM acts as a proxy for the original authors of the DI and BHC sections by generating authoring guidelines and providing topic guidance. For simplicity, topic guidance is provided indirectly and non-interactively, without refining outputs to match the target text, establishing a lower-bound baseline for performance. This simplified prompting strategy, further detailed in Appendix B, is designed to minimize user contribution and isolate the model's raw capabilities, thus providing a conservative estimate of performance in real-world interactive scenarios where iterative refinement would occur.

4.3 Human Evaluation

The evaluation of interactive, user-controlled models would ideally involve a user study, where users engage with the models to generate DI and BHC sections. However, conducting such a study at scale is beyond the scope of this exploratory study, as it is too resource-intensive and time-consuming. While comprehensive interactive studies involving clinicians actively guiding text generation are a critical next step and deferred to future work, our current preliminary human evaluations aim to demonstrate the potential of our approach and validate the LLM-as-proxy concept. We therefore complement our automatic evaluation with two preliminary human evaluations to assess the effectiveness of our approach and the quality of the dataset augmentations.

The first evaluation assessed whether our models generate clinically appropriate outputs when provided with human-written guidelines and whether automatic evaluation metrics align with human judgment. An advanced medical student in his final clinical year, serving as a domain expert, dedicated 95 hours and 13 minutes to manually authoring 600 guidelines for the DI and BHC sections of 300 randomly sampled discharge summaries from the test-phase-2 split of the 'DischargeMe!' dataset. While no fixed template was imposed, the expert was encouraged to consider elements such as document type, content coverage, structure, formatting, tone, use of language, complexity, and technicality. To ensure the guidelines captured clinically relevant stylistic and structural directives, while authentically reflecting human-written guidelines, the expert was instructed to: (1) Write naturally, following personal preferences, rather than adhering to rigid templates. (2) Provide guidance enabling a non-medical layman to write the target text solely based on the discharge summary. (3) Avoid medical jargon and patient-specific details, while capturing key clinical writing conventions.

The second evaluation assessed the quality of LLM-generated topic segmentations, specifically the topic accuracy, question validity, and text block appropriateness. 500 discharge summaries were sampled from the post-processed subset of the training split of the 'DischargeMe!' dataset for this purpose, yielding a total of 1000 segmentations $seg(t_i)$. For each t_i, one segment $[h_i^j, q_i^j, t_i^j]$ was randomly selected for assessment. The same medical expert then dedicated 26 hours and 27 minutes to evaluating each segment through a two-step process (see Appendix E).

4.3.1 Results

We evaluate BASE W/INSTR on a sample of 300 discharge summaries using human-written authoring guidelines (cf. Tab. 5). For cross-validation, we also assess BASE and BASE W/INSTR using au-

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

tomated evaluation (Sec. 4.2). Results (cf. Tab. 5) indicate that the sampled dataset is not significantly easier than the original test set (BASE: 0.369, BASE w/INSTR: 0.423) . When prompted with human-written guidelines, BASE w/INSTR (0.391) retained 92.4% of its performance, while maintaining factual consistency (AlignScore: +0.2%). This high retention rate suggests that LLM-generated guidelines align well with human-authored ones and that LLMs can serve as effective and reliable proxies for complex clinical annotation tasks like generating authoring guidelines, underscoring the promise of adopting LLMs for expert annotations in clinical practice. This outcome strongly supports the viability of using LLMs as reliable proxies for human experts in generating such guidelines.

Evaluating the topic segmentations (cf. Appendix E) reveals that 91.9% of LLM-generated headings ($\mathring{h}_i^j$) correctly match their corresponding text blocks (t_i^j). Similarly, 88.4% of the generated questions ($\mathring{q}_i^j$) are appropriately answered by the text block and effectively inquire it's content. These figures demonstrate that the LLM-generated segmentations are clinically meaningful and reliable. The selected text range (t_i^j) is deemed accurate in 75.2% of cases, meaning it accurately aligns with the optimal segment boundaries for the suggested heading $\mathring{h}_i^j$ and question $\mathring{q}_i^j$. These results suggest that expert-level accuracy may already be within reach with stronger models or a secondary validation pass to refine the segmentation, in particular segment boundaries. These findings further indicate that LLMs can reliably perform fine-grained clinical text annotation tasks typically requiring human expertise.

Conclusion: Our findings reinforce the idea that LLMs can effectively act as human proxies for complex clinical annotation tasks such as authoring guideline generation and topic segmentation, bringing automation closer to expert-level performance and enabling scalable dataset augmentation.

5 Results and Discussion

In this section, we analyze the impact of our data augmentation strategies on general instruction-tuned LLMs, evaluate the efficiency of our state-of-the-art training approach, and assess how conditioning text generation for user control enhances clinical text generation. We further present human evaluation results, validating the effectiveness of our approach.

	0-shot	3-shot	RI
Llama 3.3 70B Instruct	0.175	**0.210**	+20%
w/STYLE	0.184	**0.215**	+17%
w/INSTR	0.210	**0.223**	+6%
w/TOPICS	0.226	**0.227**	+0%
w/STYLE w/TOPICS	**0.225**	**0.225**	+0%
w/INSTR w/TOPICS	**0.230**	**0.230**	+0%

Table 1: Overall evaluation results of Llama 3.3 70B Instruct with zero-shot and three-shot prompting. Relative Improvements (RI) are rounded to integers. See Tab. 7 for detailed results.

5.1 Impact of Data Augmentations on General Instruction-Tuned LLMs

Llama 3.3 70B Instruct performs significantly worse than previous submissions on the DischargeMe! leaderboard (cf. Tab. 1 vs. Tab. 2). Nonetheless, augmenting the input with authoring guidelines and topic guidance yields a 31% relative improvement in the best configuration ($c = \texttt{topics}, g = \texttt{instr}$) over using no data augmentations. Notably, zero-shot Llama 3.3 70B Instruct w/INSTR performs on par with the three-shot setting without any dataset augmentations. This suggest that in-context learning effects can be replicated using explicit authoring guidelines with a lot less context tokens (cf. Tab. 8). Further supporting this, three-shot prompting provides no additional gains over zero-shot prompting when topic guidance is provided (w/TOPICS).

Conclusion: Overall, zero-shot Llama 3.3 70B Instruct achieves only about half of the performance of our models based on Llama 3 8B Instruct (cf. Tab.2), underscoring the importance of augmenting datasets for user control during training.

5.2 State-of-the-Art Performance with Efficient Training

Our instruction-tuned baseline model (BASE), trained without dataset augmentations, achieves a new state-of-the-art on the BioNLP ACL'24 *DischargeMe!* leaderboard, outperforming prior submissions across all metrics except METEOR (Tab. 2). BASE achieves a score of 0.363, surpassing WisPerMed (0.332), HarmonAiLab@Yale (0.300), and aehrc (0.297) — a relative improvement of 9% over the previous best model. Compared to them, BASE is more efficient:

Smaller Trainable Parameter Size. BASE has only 169M trainable parameters, which is 5-6× fewer than WisPerMed (1046M), Yale (812M), and

	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
BASE	0.363	0.168	0.483	0.255	0.345	0.472	0.362	0.359	0.460
w/STYLE	0.399	0.202	0.526	0.289	0.382	0.508	0.404	0.383	0.495
w/INSTR	0.420	0.224	0.547	0.310	0.404	0.527	0.428	0.403	0.515
w/TOPICS	0.403	0.195	0.524	0.287	0.384	0.503	0.414	0.402	0.517
w/STYLE w/TOPICS	0.436	0.226	0.562	0.319	0.421	0.539	0.444	<u>**0.429**</u>	0.548
w/INSTR w/TOPICS	<u>**0.445**</u>	<u>**0.238**</u>	<u>**0.571**</u>	<u>**0.327**</u>	<u>**0.429**</u>	<u>**0.548**</u>	<u>**0.463**</u>	0.426	<u>**0.556**</u>

Table 2: Evaluation results of the *Discharge Me!* leaderboard leaders WisPerMed (Damm et al., 2024), HarmonAiLab@Yale (Socrates et al., 2024) and aehrc (Liu et al., 2024), and our instruction-tuned models on the test set (phase 2). Bold indicates best scores in each block. In addition, underscoring indicates the overall best score. Figure 4 shows relative improvements. Table 4 breaks down performance by task. Abbreviations: BERTScore (BS), AlignScore (AS).

aehrc (894M).

Simpler Methodology. We instruction-tune Llama 3 8B Instruct, while WisPerMed employs an ensemble of instruction-tuned Llama 3 8B & 70B Instruct, OpenBioLLM 70B, Mistral 7B Instruct (v0.2), and Phi 3 Mini 128K Instruct. Yale uses an extended training dataset, while aehrc optimizes the clinical input context for downstream tasks. In addition, other submissions use nucleus sampling or 4-beam search, while we decode greedly.

Lower Computational Cost. Considering all individual training setups, our training requires only 56% of Yale's compute budget, 23% of aehrc's, and 32% of WisPerMed's.

Notably, WisPerMed and aehrc also instruction-tuned Llama 3 8B Instruct using similar approaches, yet reported significantly lower scores (0.253 and 0.235, respectively). Our model achieves relative improvements of 43% over WisPerMed's and 54% over aehrc's fine-tuning attempts. A detailed comparative analysis (Appendix C) suggests that our superior performance stems from more efficient training, which includes higher learning rates, rank-stabilized LoRA and SVD-based PISSA.

Conclusion: Our findings demonstrate that more efficient training strategies can yield substantial improvements, even with fewer parameters and lower computational costs, achieving a new state-of-the-art for clinical text generation.

5.3 Conditioning Text Generation for User Control

Authoring Guidelines. Augmenting datasets with authoring guidelines significantly improves model performance (cf. Table 2, Fig. 4). BASE w/STYLE (0.399, +10%) and BASE w/INSTR (0.420, +16%) outperform BASE (0.363), demonstrating the potential of augmenting datasets with explicit guidelines.

Style guidelines enhance lexical similarity (BLEU, ROUGE, METEOR) with 9–21% relative improvements, compensating for BASE's METEOR deficit. Surprisingly, even semantic and fact-based metrics (BERTScore, AlignScore, MEDCON) improve by 7–8%, suggesting either (i) these metrics are sensitive to stylistic variances or (ii) automatically generated style guidelines contain spurious features that reinforce factual and clinical alignment — an area requiring further research. The improvements in AlignScore are particularly noteworthy as they indicate enhanced factual consistency and a reduction in hallucinations.

Writing instructions consistently outperform style guidelines and match BASE w/TOPICS on fact-based metrics (AlignScore, MEDCON: +12%), despite the latter being conditioned for and provided with topic-level guidance.

Topic Guidance. Providing LLMs conditioned for topic-level control with topic guidance yields overall improvements (+11%) similar to style guidelines, but with fact-based metrics (AlignScore, MEDCON: +12%) contributing more. Although BASE w/STYLE w/TOPICS (+20%) performs slightly worse, integrating both authoring guidelines and topic guidance yields further performance gains across all metrics, showing that these strategies are complementary, and evidencing the need for both style- and content-aware conditioning. Notably, our best model BASE w/INSTR w/TOPICS (+22%) excels in DI generation, achieving high ROUGE-1 (0.612), BERTScore (0.587),

	Overall	BLEU	R-1	R-2	R-L	BS	METEOR	AS	MEDCON
BASE	0.119	0.008	0.201	0.040	0.126	0.174	0.132	0.107	0.161
w/STYLE	0.186	0.026	0.275	0.087	0.169	0.256	0.215	0.253	0.208
w/INSTR	0.240	0.035	0.330	0.112	0.200	0.340	0.263	0.353	0.291
w/TOPICS	0.194	0.024	0.263	0.096	0.172	0.182	0.178	**0.378**	0.260
w/STYLE w/TOPICS	0.265	0.046	0.347	0.142	0.215	0.378	0.300	0.344	0.347
w/INSTR w/TOPICS	**0.279**	**0.049**	**0.365**	**0.146**	**0.227**	**0.398**	**0.324**	0.359	**0.362**

Table 3: Evaluation of our instruction-tuned models on PubMed. Bold indicates best scores.

and MEDCON (0.594) scores. (cf. Table 4).

Conclusion: Overall, we observe that stylistic and content-related guidance is complementary, and that all metrics, even fact-based ones, appear sensitive to stylistic deviations to different degrees. Furthermore, clear instructions, expressing the purpose of stylistic features and the document clearly outperform simple stylistic descriptions.

5.4 Cross-Domain Generalization on Unseen Tasks

To assess the broader applicability of our dataset augmentation strategies, we evaluated our models, originally trained on the *DischargeMe!* dataset, on an unseen out-of-distribution biomedical task: PubMed abstract writing (Cohan et al., 2018). For this evaluation, we generated authoring guidelines and topic headings for 5000 articles sampled from the PubMed dataset's validation and test split using the same automated procedure employed for the *DischargeMe!* dataset. The PubMed dataset was used exclusively for evaluation, and both the generated guidelines and topic headings are out-of-distribution with respect to the original training data of our models.

The results (Table 3) demonstrate that our augmentation strategies substantially improve model generalization to unseen tasks. The best performing model (BASE w/INSTR w/TOPICS) achieved an average score of 0.279 on the PubMed dataset. This represents a relative improvement of approximately +134% over the BASE model (0.119) trained without any augmentations.

Conclusion: These findings indicate that conditioning models on such dataset augmentations can help models generalize more effectively across different tasks and data distributions, even those not encountered during training.

6 Conclusion and Future Work

In this work, we explored strategies for conditioning LLMs to give clinicians control over both content and style in clinical text generation. Using the BioNLP ACL'24 Shared Task Discharge Me! as a case study, we demonstrated that augmenting datasets with authoring guidelines and topic segmentation significantly improves accuracy, relevance, and factual consistency. Notably, our findings raise concerns about metrics exhibiting significant sensitivity to stylistic deviations, even when fact-based, warranting further research.

Our preliminary human evaluation suggests that LLMs can serve as proxies for expert annotations, enabling dataset augmentation at scale. By introducing a separation of content and style, we extended the traditional clinical text generation paradigm to facilitate the integration of clinical communication and authoring guidelines. Since such guidelines are crafted once per task, they offer a low-cost enhancement to clinical text generation without adding cognitive burden.

We also establish a new state-of-the-art for conventional clinical text generation on Discharge Me!, surpassing prior submissions while using fewer parameters and significantly lower computational costs. To support further research and real-world adoption, we disclose our methods, allowing hospitals and clinical institutions to adapt these augmentations to their own data and workflows.

While preliminary human evaluation validates the effectiveness of our approach, a systematic study is needed to identify which specific components of authoring guidelines contribute most to downstream performance. Future work should also focus on scaling human evaluation, assessing generalization across diverse clinical datasets, and refining LLM conditioning techniques to improve adaptability to real-world medical documentation workflows. Additionally, user studies should evaluate interactivity and its impact on clinician oversight, including detailed interactive evaluations where clinicians actively guide the generation process.

7 Limitations

While our approach demonstrates strong performance in clinical text generation, several limitations remain. Our findings rely primarily on automated metrics, with only preliminary human evaluation, making a larger-scale, clinician-in-the-loop assessment essential to validate practical usability and real-world adoption. Additionally, this study does not yet evaluate how interactive clinician involvement impacts cognitive workload and oversight burden. Future work should investigate whether LLM-conditioned generation can reduce verification effort and how user feedback can further refine dataset augmentation to better align with clinical workflows. Lastly, as with all large-scale pre-trained models, our approach inherits biases from its training data, potentially affecting fairness and reliability in clinical decision-making. No work was done to mitigate such bias and assess the clinical implications of these biases to ensure responsible AI deployment in healthcare, and the effects remain unknown.

8 Acknowledgements

This work was partly funded by the German State Ministry of Culture and Science NRW for research under the Cancer Research Center Cologne Essen (CCCE) foundation. The funding was not provided specifically for this project. We acknowledge the REACT-EU project **KITE** (Plattform für KI-Translation Essen, EFRE-0801977) for providing the computational resources used in this work. All training and evaluation were conducted on the KITE GPU cluster.

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).

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

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).

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*.

Sunjun Kweon, Junu Kim, Jiyoun Kim, Sujeong Im, Eunbyeol Cho, Seongsu Bae, Jungwoo Oh, Gyubok Lee, Jong Hak Moon, Seng Chan You, Seungjin Baek, Chang Hoon Han, Yoon Bin Jung, Yohan Jo, and Edward Choi. 2024. Publicly shareable clinical large language model built on synthetic clinical notes. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 5148–5168, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

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.

Stephanie E. Pollard, Pamela M. Neri, Allison R. Wilcox, Lynn A. Volk, Deborah H. Williams, Gordon D. Schiff, Harley Z. Ramelson, and David W. Bates. 2013. How physicians document outpatient visit notes in an electronic health record. *International Journal of Medical Informatics*, 82(1):39–46.

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.

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.

Wen Xiao, Iz Beltagy, Giuseppe Carenini, and Arman Cohan. 2022. PRIMERA: Pyramid-based masked sentence pre-training for multi-document summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5245–5263, Dublin, Ireland. Association for Computational Linguistics.

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.

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 Training Details

We fine-tune *Llama 3 8B Instruct* on 8 H100 GPUs for 3,000 steps ($\approx$ 2.8 epochs) with the AdamW 8-bit optimizer (Dettmers et al., 2022) ($\beta = (0.9, 0.999), \epsilon = 1e^{-8}$) and a batch size of 128 on completions only. 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 $\alpha_{\text{LoRA}} = 64$,

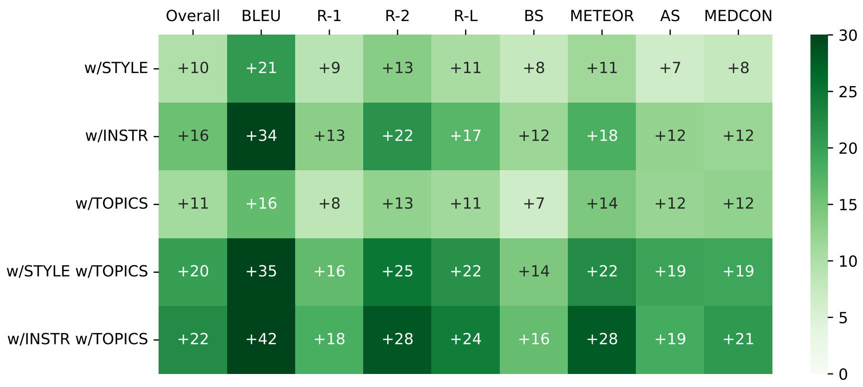

Figure 4: Relative improvement of augmented models against the traditionally instruction-tuned BASE model (cf. Tab. 2).

		Overall	BLEU	R-1	R-2	R-L	BS	Meteor	AS	MEDCON
BHC	BASE	0.333	0.142	0.465	0.228	0.313	0.460	0.335	0.290	0.435
	w/STYLE	0.350	0.158	0.488	0.242	0.330	0.477	0.356	0.297	0.452
	w/INSTR	0.364	0.171	0.505	0.255	0.343	0.489	0.376	0.304	0.470
	w/TOPICS	0.356	0.149	0.487	0.239	0.333	0.471	0.369	0.317	0.482
	w/STYLE w/TOPICS	0.385	0.178	0.524	0.268	0.367	0.504	0.396	**0.332**	0.513
	w/INSTR w/TOPICS	**0.390**	**0.183**	**0.530**	**0.271**	**0.370**	**0.509**	**0.410**	0.327	**0.517**
DI	BASE	0.393	0.193	0.502	0.283	0.377	0.484	0.390	0.428	0.484
	w/STYLE	0.448	0.247	0.565	0.337	0.434	0.539	0.452	0.469	0.538
	w/INSTR	0.476	0.277	0.589	0.366	0.464	0.565	0.480	0.503	0.560
	w/TOPICS	0.451	0.240	0.561	0.336	0.436	0.535	0.459	0.487	0.552
	w/STYLE w/TOPICS	0.487	0.273	0.600	0.370	0.474	0.574	0.491	**0.526**	0.584
	w/INSTR w/TOPICS	**0.500**	**0.292**	**0.612**	**0.383**	**0.488**	**0.587**	**0.517**	0.525	**0.594**

Table 4: The average scores per metric of our evaluation (Sec. 5.3), broken down by the two tasks: discharge instructions (DI) generation and brief hospital course (BHC) generation.

	Overall	BLEU	R-1	R-2	R-L	BS	METEOR	AS	MEDCON
BASE	0.369	0.175	0.488	0.261	0.355	0.480	0.369	0.362	0.460
w/INSTR (A)	**0.423**	**0.228**	**0.551**	**0.314**	**0.411**	**0.531**	**0.435**	0.403	**0.514**
w/INSTR (H)	0.391	0.183	0.511	0.278	0.379	0.499	0.385	**0.404**	0.491

Table 5: The results of BASE and BASE w/INSTR evaluated on 300 discharge summaries randomly sampled from the *Discharge Me!* test phase 2 split, once with augmented authoring guidelines (A) and once with human-written authoring guidelines (H).

$\text{dropout}_{\text{LoRA}} = 0.1$, $r_{\text{LoRA}} = 64$, 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 $\alpha_3 = 1e^{-6}$. This increases the duration of high learning rates even further.

B Automated Evaluation

For evaluation, we use the code provided by the *Discharge Me!* challenge, which employs a comprehensive set of metrics to assess lexical and semantic similarity, factual consistency, as well as the clinical relevance and correctness.

The metrics include: **BLEU-4** (Papineni et al., 2002), which 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), which 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), which uses contextual embeddings from BERT to evaluate the semantic similarity between the generated and reference texts. **Meteor** (Banerjee and Lavie, 2005), which considers synonyms and stemming to compare the generated text with reference texts, providing a more flexible measure of similarity. **AlignScore** (Zha et al., 2023), which 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), which is specifically designed for medical contexts, and evaluates the clinical relevance and correctness of the generated text.

To simulate user control for automated evaluation on the *DischargeMe!* dataset, we employ Llama 3.1 70B Instruct again to automatically generate authoring guidelines (Sec. 3.3) and provide topic-level control. The LLM serves as proxy for the original authors based on the assumption that their generated output approximates the input and feedback the authors would have provided if they had originally used our methods to write the target texts t_i.

While authoring guidelines can be seamlessly incorporated at inference time for **W/STYLE** and **W/INSTR**, simulating granular, interactive topic-level control for **W/TOPICS**, which iteratively refines model output, is more complex. Although increased user interaction generally improves output quality, it also amplifies user contribution, making results less indicative of the model's standalone performance. To minimize this, we provide topic guidance indirectly and non-interactively, effectively establishing a lower-bound baseline. This simplified prompting strategy strengthens our findings by isolating the model's raw capabilities while limiting user influence. As such, it offers a conservative estimate of system performance in real-world interactive settings, capturing the benefit of guided input while remaining independent of user skill or intervention.

Specifically, we extend the user prompt $user_i(c, g)$ with an instruction to cover a predefined list of topics (cf. Fig. 5). This list is derived from topic segmentations (Sec. 3.2) for each target text t_i by extracting and concatenating the headings $\mathring{h}_i^1$, ..., $\mathring{h}_i^n$ into an unnumbered bullet list.

User Message

```
{{discharge summary}}
{{radiology report 1}}
        ...
{{radiology report n}}
{{brief hospital course}}
{{authoring guidelines}}
{{instructions}}
{{topics}}
```

Figure 5: The user prompt used for evaluation.

C Comparative Analysis with Existing Approaches

We present a detailed comparison of our instruction-tuned Llama 3 8B Instruct BASE model against the three top-performing systems on the *DischargeMe!* leaderboard. We also include a detailed comparison of other instruction-tuned Llama 3 8B Instruct models evaluated during experimentation by leaderboard participants but ultimately dropped due to suboptimal performance. Table 6 summarizes the primary distinctions across these methods.

WisPerMed (Damm et al., 2024) achieved the highest leaderboard score of 0.332, surpassing other submissions by a notable margin. This success was attributed to its Dynamic Expert Selection (DES) strategy, which combines predictions from five instruction-tuned models: Llama 3 8B and 70B Instruct, OpenBioLLM 70B, Mistral 7B Instruct (v0.2), and Phi 3 Mini 128K Instruct. Notably, the standalone Llama 3 8B Instruct model within this ensemble achieved the lowest score (0.253), marginally underperforming the Phi 3 Mini model.

All models in WisPerMed's ensemble were fine-tuned using the entire discharge summary as input and LoRA with a rank of $r_{LoRA} = 16$, applied to all linear layers. In addition, some models were fine-tuned on Asclepius (Kweon et al., 2024). For inference, they employed optimized nucleus sampling to enhance output quality. This ensemble approach enabled WisPerMed to leverage complementary model strengths, albeit at the cost of increased complexity and resource demands.

aehrc (Liu et al., 2024), similarly, fine-tuned the Llama 3 8B Instruct model using LoRA ($r_{\mathrm{LoRA}} = 64$), but introduced notable variations in preprocessing and decoding strategies. Discharge summaries were partitioned into: (1) the text preceding the Brief Hospital Course (BHC) section for BHC generation, and (2) the text between the BHC and Discharge Instructions (DI) sections, joined

	BASE (ours)	WisPerMed	aehrc
Score	<u>**0.363**</u>	**0.253**	**0.235**
Clinical Context	$ds + rr$	ds	optimized
Models trained	1	2	2
Decoding Strategy	greedy	optimized nucleus sampling	4-beam search
Optimizer	AdamW 8-Bit	AdamW 8-Bit	Adam
Epochs	2.8	3	5
Batch Size	128	16	16
Learning Rate α	$1e^{-4}$	$2e^{-4}$	$2e^{-4}$
Weight Decay	$1e^{-4}$	$1e^{-2}$	N/A
Learning Rate Scheduler	optimized WSD	linear	linear
Warmup Steps	0	5	3%
LoRA Type	rank-stabilized LoRA	LoRA	QLoRA
Layers	all linear	all linear	all linear
Total Trainable Parameters	168M	84M	336M
Weight Initialization	fast SVD-PISSA ($n = 32$)	N/A	QLoRA
r_{LoRA}	64	16	64
α_{LoRA}	64	16	16
dropout$_{LoRA}$	0.1	0	N/A

Table 6: Detailed comparison of training configurations, decoding strategies, and scores for instruction-tuned Llama 3 8B Instruct models, highlighting the key differences among our, WisPerMed's and aehrc's approach. ds = Discharge Summary. rr = Radiology Reports. N/A = Not Available.

with the BHC section, for DI generation. This design was motivated by their observation that longer input contexts negatively impacted model performance. They also reported that providing the entire discharge summary, including all radiology reports (as used in our setting), yielded the lowest results. For decoding, aehrc employed a 4-beam search strategy. Their leaderboard submission leveraged PRIMERA (Xiao et al., 2022), a specialized instruction-tuned summarization model with 447M parameters.

HarmonAiLab@Yale (Socrates et al., 2024) has not experimented with Llama 3 8B Instruct, but GPT-3 and GPT-4 instead. They ultimately submitted a fine-tuned clinical model (BioBART-Large, 406M parameters) trained on an extended dataset that reportedly included samples from the validation and phase 1 test splits for a total of 83.475 (+21.4%) training samples for the BHC task. HarmonAiLab@Yale also employed a 4-beam search strategy for generation, but blocking repeats with an n-gram size of 3.

All teams (WisPerMed, aehrc, and HarmonAiLab@Yale) trained separate models for BHC and DI tasks, effectively doubling their total trainable parameter size.

In contrast, we adopt a unified strategy, training a single Llama 3 8B Instruct model to jointly handle both BHC and DI tasks. The input includes the entire discharge summary and all radiology reports. Similar to aehrc, we applied LoRA ($r_{\text{LoRA}} = 64$) to all linear layers during fine-tuning, resulting in a total trainable parameter count of 168M – double that of WisPerMed but only half that of aehrc when comparing the fine-tuned Llama 3 8B Instruct models (rather than final submissions). For decoding, we employed a greedy decoding strategy. See Table 6 for a more detailed comparison.

Despite the less favorable input context and decoding strategy, our model achieved a leaderboard score of 0.363 — a 43% improvement over WisPerMed's attempts, and a 54% improvement over aehrc's attempts with instruction-tuned Llama 3 8B Instruct models (cf. Tab. 6). Moreover, our method (168M) has significantly fewer total trainable parameters than the final submissions of WisPerMed (1.046B), HarmonAiLab@Yale's (812M) and aehrc's (894M), requires less training (2.8 epochs) than WisPerMed (3 epochs) and aehrc (5 epochs), and no additional data (WisPerMed), nor an extended dataset (HarmonAiLab@Yale). Considering all individual training setups, our training also requires only 56% of Yale's compute budget, 23% of aehrc's, and 32% of WisPerMed's

The results underscore the efficiency and effectiveness of our approach, demonstrating that instruction-tuning a single general-purpose model can achieve state-of-the-art performance without the complexity of ensembles or reliance on domain-specific models and architectures.

D Topic Segmentation Post-Processing

This step is applied only when the segmentation introduces minor alterations to the original text to avoid introducing inconsistencies between headings, questions, and text blocks through such replacements. To achieve this, we use diff methods to identify word-level differences — defined as whitespace-delimited character sequences — between the generated text $\mathring{t}_i = (\mathring{t}_i^1, \dots, \mathring{t}_i^n)$ and original text t_i. Segmentations containing consecutive differences are then filtered out, ensuring that segmentations involving only minor differences, such as the spelling or formatting, are considering for this post-processing step. This leaves us with 93.61% of the DI, and 81.15% of the BHC segmentations, whose blocks t_i^k are then mapped back to the original text t_i to retrieve the original character sequences corresponding to each block.

E Human Evaluation of Topic Segmentations

We conducted a human evaluation of the topic segmentations (Sec. 3.2) generated using Llama 3.1 70B Instruct. Specifically, 500 DI and BHC sections were randomly sampled from the post-processed subset of the training split of the *DischargeMe!* dataset, resulting in a total of 1000 target texts t_i. For each t_i, one segment $seg_i^j = (h_i^j, q_i^j, t_i^j)$ was randomly selection for assessment. A human expert then evaluated the selected segment through a two-step process, details in Section E.

Step 1. The expert was presented with the target text t_i, where the text range from the start of the selected text block t_i^j to the end of t_i was highlighted. The expert was instructed to annotate the next topic beginning within the highlighted range by identifying the heading, question, and corresponding text block. This step ensured that the expert interacted thoroughly with the target text and independently assessed and annotated the next segment without being influenced by the LLM-generated output.

Step 2. The expert was then provided with the LLM-generated annotation $s\mathring{e}g_i^j$, which included the heading $\mathring{h}_i^j$, question $\mathring{q}_i^j$, and text block t_i^j. The expert evaluated the appropriateness of the generated heading, the quality of the question, and the accuracy of the text block boundaries. While the expert could refer to their own annotations for comparison, they were instructed to assess the LLM-generated segment for correctness without imposing personal preferences, given the inherent subjectivity of the topic segmentation task and the existence of multiple competing solutions.

The heading $\mathring{h}_i^j$ was considered appropriate only if it effectively encapsulated the content and focus of the corresponding text block t_i^j. The question $\mathring{q}_i^j$ was considered high quality only if it was directly answerable by the selected text block t_i^j and accurately reflected and inquired the central issue or information addressed within that range.

For evaluating the text range, the expert was tasked with envisioning the optimal segment boundaries, aligning with both the heading $\mathring{h}_i^j$ and the Question Under Discussion (QUD) $\mathring{q}_i^j$, within the entire target text t_i. The text range was considered accurate only when the start and end points coincided with the hypothesized segment boundaries.

Results The evaluation revealed that the LLM-generated headings ($\mathring{h}_i^j$) aligned with the corresponding text blocks (t_i^j) in the majority of cases (91.9%). Similarly, the generated questions ($\mathring{q}_i^j$) were well-formulated in 88.4% of instances, effectively inquiring about the content of the text block and being answerable by it. In 87.5% cases, both the heading and question was deemed appropriate. The accuracy of the selected text range (t_i^j) was confirmed in 75.2% of all cases. Notably, in instances where both the headings and questions were appropriate, the accuracy of the text ranges increased to 80.91%.

F Dataset & Annotation Statistics, Prompts and Examples

In this section, we present examples of data augmentations, showcasing annotation samples alongside corresponding LLM prompts. Table 8 summarizes token length statistics for the DischargeMe! training split. We find that style guidelines and writing instructions have similar average lengths across tasks (DI vs. BHC), but writing instructions are nearly 1.5× longer than style guidelines. Additionally, BHC sections are, on average, twice as long as DI sections, and BHC topic segmentations consistently contain slightly more segments, longer headings, and extended text blocks, as detailed in

	Overall	BLEU	R-1	R-2	R-L	BS	Meteor	AS	MEDCON
Llama 3.3 70B Instruct (0-shot)	0.175	0.012	0.286	0.069	0.142	0.209	0.205	0.242	0.239
w/STYLE	0.184	0.014	0.300	0.070	0.148	0.231	0.220	0.243	0.246
w/INSTR	0.210	0.021	0.333	0.087	0.170	0.276	0.246	0.269	0.279
w/TOPICS	0.226	0.027	0.352	0.095	0.183	0.292	0.283	0.260	0.320
w/STYLE w/TOPICS	0.225	0.027	0.348	0.094	0.182	0.291	0.281	0.259	0.319
w/INSTR w/TOPICS	0.230	0.027	0.352	0.097	0.185	0.296	0.287	0.270	0.322
Llama 3.3 70B Instruct (3-shot)	0.210	0.025	0.336	0.094	0.177	0.281	0.232	0.246	0.285
w/STYLE	0.215	0.027	0.342	0.097	0.183	0.296	0.235	0.254	0.289
w/INSTR	0.223	0.029	0.345	0.104	0.193	0.317	0.237	0.259	0.299
w/TOPICS	0.227	0.027	0.352	0.095	0.183	0.292	0.283	0.261	0.321
w/STYLE w/TOPICS	0.225	0.027	0.348	0.094	0.182	0.291	0.281	0.259	0.319
w/INSTR w/TOPICS	0.230	0.028	0.352	0.097	0.185	0.296	0.287	0.271	0.322

Table 7: The average scores per metrics for our evaluations, broken down by the two tasks: discharge instructions (DI) generation and brief hospital course (BHC) generation.

	DS + RRs	DI	BHC
`#tokens`	3883.76 (± 2262.69)	278.68 (± 220.91)	525.05 (± 386.98)
`#tokens(SG)`	—	330.62 (± 34.51)	325.54 (± 29.73)
`#tokens(WI)`	—	470.58 (± 42.21)	460.46 (± 37.19)

Table 8: Averages (and standard deviations) of token counts for various quantities of the augmented DischargeMe! training split. Abbreviations: SG = Style Guidelines. WI = Writing Instructions. DS = Discharge Summary. RRs = Radiology Reports. DI = Discharge Instructions. BHC = Brief Hospital Course.

	DI	BHC
`#segments`	6.25 (± 2.09)	8.25 (± 4.03)
`#tokens(`$\mathring{h}_i^k$`)`	3.82 (± 2.00)	4.65 (± 2.95)
`#tokens(`$\mathring{q}_i^k$`)`	9.60 (± 2.64)	11.46 (± 2.97)
`#tokens(`t_i^k`)`	44.57 (± 46.16)	62.51 (± 61.43)

Table 9: Averages (and standard deviations) of various quantities of topic segmentations for DI and BHC sections of the DischargeMe! training split. The statistics for the number of segments `#segments` and the token counts `#tokens(·)` of headings $\mathring{h}_i^k$, questions $\mathring{q}_i^k$ and text blocks t_i^k are consistently larger for BHC sections.

Table 9.

Synthetic Clinical Document
The 75-year-old male patient with multi-organ sarcoidosis, diabetes mellitus, and chronic renal failure was admitted due to fatigue, dyspnea, lower limb edema, and pain. He received corticosteroid therapy for two years but experienced a bloodstream infection caused by Pseudomonas aeruginosa, which was successfully treated with levofloxacin. The patient’s dosage of methylprednisolone was increased, leading to him being transferred to ICU and intubated due to worsening functional status. He was diagnosed with Candida albicans on Day +3 and started antifungal therapy with fluconazole (400 mg daily) and then later found to have disseminated cryptococcal disease on Day +5, leading to antifungal therapy with liposomal amphotericin B (80 mg daily). Unfortunately, the patient died from septic shock on Day +10. The laboratory findings indicated lymphocytopenia of 900 cells/μL, creatinine of 1.73 mg/dL, C-reactive protein of 83 mg/L, procalcitonin of 2.5 ng/L, increased C-reactive protein to 160 mg/L, increased procalcitonin to 14 ng/mL, and serum positive titers for CrAg ($\geq 1 : 4096$). The diagnostic findings included pulmonary infiltration with lymphadenopathy, multiple nodules within the lung parenchyma, and disseminated cryptococcal disease. The treatment consisted of broad-spectrum antibiotic therapy with meropenem and teicoplanin, antifungal therapy with fluconazole (Day +3), and antifungal therapy with liposomal amphotericin B (Day +6). There is no follow-up information available.
Topic Segmentation Prompt
You are tasked with fine-grained topic segmentation. Given this formatted text, segment the paragraphs into as many short blocks as sensible, each with a distinct topic. Give each block a meaningful, short topic heading, summarizing the most important information from the beginning of the block for the intended audience, and a subtitle, which reformulates the topic as a question that is answered by the block. Guidelines: - Segment everything from the very first to the very last word/character/symbol. - Terminate spans and insert new headings, whenever the upcoming text does not match the current running topic anymore, e.g. whenever the medical, clinical or healthcare focus changes. - When formulating questions, do not use pronouns as the subjects, and do not use possessive pronouns. - Do not alter the text. Copy typos, errors, mistakes and formatting from the original text. - Include headings, symbols, separators, vertical/horizontal spacing, empty lines and other formattings with their associated blocks. Answer format: ’<split-text>\n<topic>...</topic>\n<question>...</question>\n<span>...</span>\n\n<topic>...</topic>\n<question>...</question>\n<span>...</span>\n\n...</split-text>’
Segmented Text
<topic>Initial Patient Information**</topic>** **<question>**What is the patient’s initial condition and medical history?**</question>** **<span>**The 75-year-old male patient with [...] fatigue, dyspnea, lower limb edema, and pain.**</span>** **<topic>**Initial Treatment and Complications**</topic>** **<question>**What treatment did the patient receive initially, and what complications arose?**</question>** **<span>**He received corticosteroid therapy for [...] intubated due to worsening functional status.**</span>** **<topic>**Antifungal Therapy and Diagnosis**</topic>** **<question>**What antifungal therapy was administered, and what diagnoses were made?**</question>** **<span>**He was diagnosed with Candida albicans [...] liposomal amphotericin B (80 mg daily).**</span>** **<topic>**Outcome and Laboratory Findings**</topic>** **<question>**What was the patient’s outcome, and what laboratory findings were reported?**</question>** **<span>**Unfortunately, the patient died from septic shock [...] positive titers for CrAg ($\geq 1 : 4096$).**</span>** **<topic>**Diagnostic Findings**</topic>** **<question>**What diagnostic findings were reported?**</question>** **<span>**The diagnostic findings included [...] parenchyma, and disseminated cryptococcal disease.**</span>** **<topic>**Treatment Summary**</topic>** **<question>**What treatment did the patient receive?**</question>** **<span>**The treatment consisted of broad-spectrum [...] liposomal amphotericin B (Day +6).**</span>** **<topic>**Follow-up Information**</topic>** **<question>**Is there any follow-up information available?**</question>** **<span>**There is no follow-up information available.**</span>**

Table 10: Topic segmentation of our framework of an arbitrary clinical document retrieved from the synthetic Asclepius dataset (Kweon et al., 2024) for demonstration purposes, as the *DischargeMe!* dataset cannot be used directly due to privacy restrictions and access limitations. The control sequences to initiate and stop generation of specific elements are indicated in bold.

Prompt for Style Guideline Generation

<text>{{target text}}</text>
Describe the text's tone, writing style, document format, layout, composition, textual structure, use of language, use of abbreviations, use of medical jargon, the intendened audience and anything else noteworthy. Write full sentences and paragraphs.

Guidelines:
- Do not use the terms from the text.
- Do not quote the text.
- Do not give examples from the text.
- Do not reveal details about the patient.

Style Guideline

The original text's tone is formal, objective, and neutral, indicating a professional and clinical approach to documenting a patient's medical history and treatment. The writing style is concise, direct, and informative, suggesting a focus on conveying complex medical information in a clear and efficient manner. The document format appears to be a medical case report or a patient summary, likely used for educational, research, or clinical purposes. The layout is structured and organized, with a logical flow of information that facilitates easy comprehension. The composition is dense and detailed, indicating a high level of technical expertise and a focus on accuracy.

The textual structure is chronological, with events and treatments presented in a linear sequence. This structure allows the reader to follow the patient's progression and the corresponding medical interventions. The use of language is technical and specialized, with a high degree of medical terminology and jargon. This suggests that the intended audience is composed of medical professionals or individuals with a strong background in healthcare.

The use of abbreviations is frequent, indicating a familiarity with medical shorthand and a desire to convey complex information concisely. The text assumes a high level of prior knowledge in medicine, with minimal explanation of technical terms or concepts. This suggests that the intended audience is composed of specialists or experts in the field.

The text's use of medical jargon and technical terminology creates a sense of authority and expertise, but may also limit its accessibility to non-medical readers. The language is precise and formal, with a focus on conveying complex information in a clear and concise manner.

Overall, the text appears to be written for an audience of medical professionals, researchers, or students in the healthcare field. The tone, writing style, and use of language all suggest a focus on conveying complex medical information in a clear and efficient manner, with a high degree of technical expertise and accuracy.

Table 11: The Style Guideline generated for the synthetic clinical document from Tab. 10, constituting of a list of descriptions of stylistic features.

Prompt for Writing Instructions Generation

<text>{{target text}}</text>
Please provide detailed and comprehensive writing instructions for a non-specialist to exactly reproduce the text above. The instructions should include details on:
- the purpose and intent of the text, including how it is achieved
- the intended audience, including how the audience's needs are met
- the tone of text, including how to achieve it
- the text structure and outline
- the text disposition
- the text formatting (not typographical), such as (but not only) the use of paragraphs, subheadings, introductions, closings, bullet points, list, including any apparent rules and patterns
- the use of language, including the use of abbreviations and medical jargon (if it is used), with respect to the audience
- and any other noteworty features.

Guidelines:
- Use an instructive tone for writing.
- Consider that the non-specialist will not see the original text.
- Do not use the terms from the text.
- Do not quote the text.
- Do not give examples from the text.
- Do not reveal details about the patient.

Answer format: '## Writing Instructions\n\n...'

Writing Instructions

## Writing Instructions

### Purpose and Intent

The purpose of this text is to document a patient's medical history, diagnosis, treatment, and outcome in a clear and concise manner. The intent is to provide a comprehensive record of the patient's case, which can be used for medical reference, education, or research purposes. To achieve this, focus on presenting the information in a logical and chronological order, using specific details and medical terminology.

### Intended Audience

The intended audience appears to be medical professionals, researchers, or students who require in-depth information about the patient's case. To meet the audience's needs, use technical language and assume a certain level of medical knowledge. However, avoid using overly complex terminology or jargon that might be unfamiliar to non-specialists in the field.

### Tone

The tone of the text is formal, objective, and informative. To achieve this tone, maintain a neutral and detached perspective, focusing on presenting the facts of the case without emotional or personal commentary. Use a professional and respectful language throughout the text.

### Text Structure and Outline

The text follows a chronological structure, starting with the patient's admission and progressing through their diagnosis, treatment, and outcome. The outline appears to be:
1. Introduction: Brief overview of the patient's condition and admission
2. Medical history: Description of the patient's pre-existing conditions and previous treatments
3. Diagnosis: Presentation of the patient's symptoms, laboratory findings, and diagnostic results
4. Treatment: Description of the treatments administered, including medications and therapies
5. Outcome: Report of the patient's outcome, including any complications or results

### Text Disposition

The text is written in a straightforward and factual style, with no apparent attempt to persuade or entertain the reader. The focus is on conveying information in a clear and concise manner.

### Text Formatting

The text appears to be written in a single block of paragraphs, with no subheadings, bullet points, or lists. The paragraphs are relatively short, with each one focusing on a specific aspect of the patient's case. To replicate this format, use a single block of text, dividing it into paragraphs as needed to separate distinct sections of information.

### Language and Terminology

The text uses medical terminology and abbreviations, assuming a certain level of familiarity with medical concepts. To replicate this, use technical language and abbreviations commonly used in medical contexts. However, avoid using overly complex or specialized terminology that might be unfamiliar to non-specialists in the field.

### Noteworthy Features

The text includes specific details about the patient's laboratory findings, medications, and treatments, which suggests a focus on accuracy and precision. To replicate this, include specific details and data wherever possible, using precise language and terminology.

Table 12: The Writing Instructions generated for the synthetic clinical document from Tab. 10 alongside the respective prompt. Writing Instructions are structured and more comprehensive than Style Guidelines. They also feature a slight instructional tone.