國立臺灣大學電機資訊學院資訊網路與多媒體研究所
碩士論文
Graduate Institute of Networking and Multimedia
College of Electrical Engineering and Computer Science
National Taiwan University
Master's Thesis

基於本地大型語言模型的門診對話臨床紀錄摘要
Clinical Note Summarization from Outpatient Conversations Using Local LLMs

陳彥錞

Yen-Chun Chen

指導教授：陳信希 博士

Advisor: Hsin-Hsi Chen, Ph.D.

中華民國 114 年 8 月

August, 2025

doi:10.6342/NTU202504251

致謝

回顧三年間的研究生活，是我求學階段最迷惘而不安的時期，最後能夠完成這份碩士論文，我衷心感謝所有給予我幫助與支持的人們。 首先，誠摯感謝我的指導教授陳信希老師，在我最爲低潮的期間，體諒我的焦慮，等待我一步步克服煩惱，慢慢摸索與學習，這三年間耐心地指導我如何思考研究的設計，並以積極的心態面對實驗結果。老師的熱情指導與勉勵，總使我獲益良多，讓我更加理解如何進行學術研究，在此致上最深的感謝。

此外，我也很感謝實驗室的各位，謝謝陳建宏學長一路上陪伴我完成研究，不僅總是幫助我完善實驗設計，還一直鼓勵我調整心態、緩解緊張；感謝吳承光學長在忙碌之餘引導我尋找研究題目，體諒我的不成熟，留給我足夠的機會重新省視自己的前進方向。同時，我也很感謝黃瀚萱學長、陳柏君學長，不時針對我的進度報告提供建議；也謝謝同屆的曾郁、林晉毅一同討論課程報告；在口試準備期間，也很謝謝張庭維、林緯翔、王睿誼的協助，讓我可以專心完成自己的論文。

最後，衷心感謝我的父母，一直以來全力支持我努力完成學業，並在生活上無微不至的照顧我，使我能夠無後顧之憂地專注於學習與研究，在我碰壁失落的時候，總是陪伴及鼓勵我繼續前進。我也非常感謝我的哥哥，即使自己也是研究生，從過去至今，總是和我分享無論是學習或調適心態的方法與知識，幫助我解決各種疑難雜症，互相討論彼此的研究，協助我釐清思路，在口試的準備上也給予許多支持。我也要感謝自己，即使充滿不安與徬徨，依然堅持不放棄，相信自己最後能完成這份研究。

doi:10.6342/NTU202504251

再次感謝所有在這段歷程中陪伴及幫助我的人們，在此獻上我最真摯的感謝與敬意，我會珍惜這一切收穫，踏實地朝未來邁進。

doi:10.6342/NTU202504251

摘要

大型語言模型近年來被視爲簡化臨床工作流程的有力工具。然而，在高度敏感的醫療領域中應用模型面臨諸多挑戰，例如隱私保護，以及缺乏公開且高品質的臨床對話資料集。本研究聚焦於使用開源大型語言模型，在完全本地環境下，從實際的門診對話中產生臨床紀錄，以確保病患隱私不外洩。我們設計了一套完整的資料前處理流程，包含對實際醫療對話的摘要與翻譯，並重新標註對應的臨床紀錄內容。本文探討三種臨床紀錄生成方式：單階段端到端生成、兩階段檢索增強生成、以及單階段生成搭配合成對話擴充。我們的實驗顯示監督式微調在效能上表現優異，且小型模型在準確檢索關鍵證據方面亦展現潛力。儘管大型語言模型可在一定程度上協助摘要臨床紀錄，但要維持完全在地部署兼顧效能仍是一大挑戰。本論文突顯了當前大型語言模型應用於醫療資料的潛力與限制，特別是在隱私要求高、需本地部署的場景下。

關鍵詞: 臨床紀錄生成、大型語言模型、醫療對話、隱私保護、本地部署

doi:10.6342/NTU202504251

Abstract

Large language models (LLMs) have emerged as a promising tool to streamline clinical workflows. However, the application of LLMs in the highly sensitive domain of healthcare faces major challenges, such as strict privacy regulations and the scarcity of publicly available, high-quality clinical dialogue datasets. This work focuses on clinical note generation from real-world outpatient conversations using open-source LLMs in a fully local environment to preserve patient privacy. We developed a comprehensive data preprocessing pipeline involving summarization and translation of real-life medical dialogues, along with meticulous re-annotation of the corresponding clinical notes. Three approaches to note generation are explored: One-stage End-to-end Generation, Two-stage Retrieval-Augmented Generation and One-stage Generation with Synthetic Dialogue Augmentation. Our experiments demonstrated the effectiveness of supervised fine-tuning methods and the the potential of smaller models in accurately retrieving evidence. While LLM applications can assist in summarizing clinical notes to a certain extent, maintaining fully local models for privacy remains a significant challenge. This work highlights both the potential and the limitations of current LLM-based approaches in this specialized domain, particularly under local deployment constraints.

Keywords: Clinical Note Generation, Large Language Models, Medical Dialogue, Privacy, Local Deployment

doi:10.6342/NTU202504251

Contents

doi:10.6342/NTU202504251

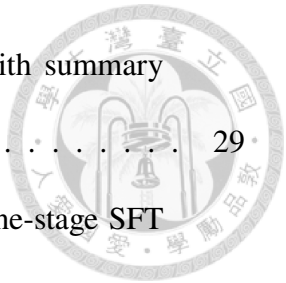
doi:10.6342/NTU202504251

List of Figures

doi:10.6342/NTU202504251

List of Tables

doi:10.6342/NTU202504251



doi:10.6342/NTU202504251

Chapter 1. Introduction

Physician—patient conversations are central to the clinical workflow, serving not only as a means for diagnosis and treatment but also as the foundation for generating critical medical documentation. In outpatient settings, doctors collect patient information, such as symptoms, family history, and lifestyle habits, through dialogue, and then summarize the key points into structured clinical notes for future reference. One widely adopted format for this documentation is the SOAP note, which includes four sections: Subjective, Objective, Assessment, and Plan.

Despite the importance of clinical documentation, it is time-consuming and requires a significant amount of manual effort, placing a heavy burden on clinicians. The documentation workload is widely recognized as a contributing factor to physician burnout [7, 8].

To alleviate this burden, automatic clinical note generation has emerged as a promising solution. Recent advances in large language models (LLMs) have demonstrated their ability to understand and generate human-like language, offering new opportunities to streamline clinical workflows. However, the application of LLMs in the highly sensitive domain of healthcare faces major challenges, such as strict privacy regulations and the scarcity of publicly available, high-quality clinical dialogue datasets.

Additionally, automating clinical note generation from real-world outpatient conversations introduces further complexity. Unlike curated online consultations, outpatient interactions are spontaneous, unstructured, and often noisy due to speech disfluencies, overlapping dialogue, and environmental interference, which introduce variability and transcription errors. Meanwhile, the target notes are often incomplete, filled with domain-specific abbreviations, or grammatically incorrect, further complicating modeling.

doi:10.6342/NTU202504251

Moreover, collecting and annotating real-world data is non-trivial. Ethical constraints, privacy concerns, and the need for informed consent restrict access to patient conversations. High-quality annotation requires medical expertise, increasing both cost and time, and limiting the scalability of dataset creation. As a result, most prior work in medical dialogue summarization relies on synthetic or publicly available online consultation datasets [3, 4], which lack the complexity and realism of real-world clinical interactions.

Motivated by these gaps, this thesis focuses on generating structured clinical notes from real-world outpatient conversations in a local-only environment to ensure privacy preservation.

Our contributions are summarized as follows:

1. We develop a robust data preprocessing pipeline, including summarization and translation of real-world medical dialogues, along with meticulous re-annotation of the corresponding clinical notes.

2. We investigate three paradigms for clinical note generation:

   (a) **One-stage End-to-end Generation**: Leveraging LLMs to directly generate clinical notes.

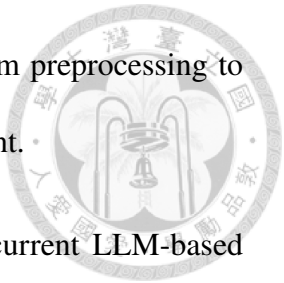
   (b) **Two-stage Retrieval-Augmented Generation**: Utilizing the retrieval capabilities of small LLMs to locate relevant information from lengthy contexts.

   (c) **One-stage Generation with Synthetic Dialogue Augmentation**: Expanding training datasets with generated dialogues to improve robustness.

3. We provide qualitative case studies to analyze the generation quality of different approaches.

doi:10.6342/NTU202504251

4. We demonstrate the feasibility of running the full pipeline(from preprocessing to generation) using only open-source LLMs in a local environment.



This work highlights both the potential and the limitations of current LLM-based approaches in this specialized domain, particularly under local deployment constraints. Our findings emphasize the need for expert evaluation and further development to enable reliable clinical applications.

doi:10.6342/NTU202504251

Chapter 2. Related Work

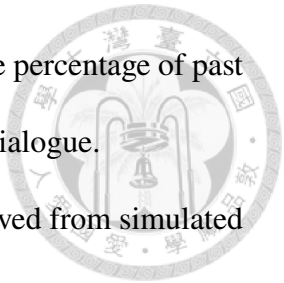
2.1 Clinical Note Generation

Automated clinical note generation has gained increasing attention as a means to reduce physicians' documentation burden and improve the efficiency of electronic health record (EHR) systems. Most studies focus on transforming doctor—patient conversations into structured notes, typically in the SOAP (Subjective, Objective, Assessment, Plan) format.

The Wang Lab [10] participated in the MEDIQA-Chat 2023 shared task and reported results for two approaches: the first fine-tuned a pre-trained language model (PLM) on the shared task data, and the second used few-shot in-context learning (ICL) with a large language model (LLM). They achieved the highest score in the shared tasks with similar-dialogue's notes as few-shot to perform an ICL-based approach using GPT-4. LLM fine-tuning methods were further investigated by Leong et al. [15]. They explored parameter-efficient fine-tuning methods for LLMs and demonstrated that even lightweight adaptations (e.g., LoRA) can achieve competitive performance while reducing computational cost. Their findings are particularly relevant for local deployment, where resource constraints are critical.

Chen et al. [12] propose two frameworks to support automatic medical consultation: doctor-patient dialogue understanding and task-oriented interaction. Key tasks include named entity recognition, dialogue act classification, symptom label inference, medical report generation, and diagnosis-oriented dialogue policy. They create IMCS-21, a large-scale annotated medical dialogue corpus, as a benchmark for medical dialogue modeling.

doi:10.6342/NTU202504251



Dialogue uncoverage issues are also mentioned, such as a considerable percentage of past medical history being empty because this part is less involved in the dialogue.

Microsoft Health AI [13] introduces the MTS-Dialog dataset, derived from simulated doctor-patient conversations based on publicly available clinical notes. This dataset is used to train transformer-based models (e.g., BART and PEGASUS) and variantss, which are pre-finetuned on relevant corpora to summarize conversations into clinical notes. Data augmentation via back-translation is employed to enhance model robustness. Section headers from the clinical notes serve as prefix signals to guide the summarization process. Pre-finetuning and signal guidance improve factual performance and summary fluency, while reducing critical fact omissions. Nonetheless, the best-performing model still exhibited a hallucination rate of 3% and failed to capture 33% of the medical facts.

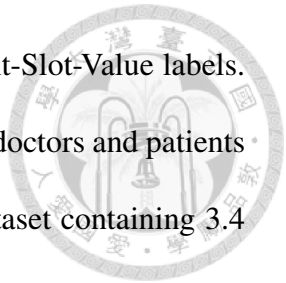
2.2 Medical Dialogue Datasets

Several works focus on constructing medical dialogue datasets. The ACI-Bench dataset [2] comprises various types of doctor-patient interactions. Some dialogues involve physicians interacting with a virtual assistant using fixed commands, while others include physicians collaborating with human or virtual scribes to compose clinical notes. The remaining data are generated through role-playing between a certified physician and a layperson volunteer, based on symptom prompts.

In addition, several large-scale datasets have been created from online medical consultation platforms in China. These datasets are particularly useful for improving performance on medical dialogue generation tasks, especially with smaller models.

ReMeDi [4] is built by crawling raw medical dialogues and medical knowledge bases from online websites. The dialogues are cleaned using a set of rigorous rules, and

doi:10.6342/NTU202504251



annotated following well-defined guidelines to produce detailed Intent-Slot-Value labels. The final dataset includes 1,557 out of 96,965 conversations between doctors and patients with fine-grained annotations. MedDialog [3] is another massive dataset containing 3.4 million doctor-patient conversations in Chinese and 0.26 million in English. Models trained on MedDialog have been shown to generate clinically correct and human-like medical dialogues.

2.3 Synthetic Data Generation

Numerous studies have explored general synthetic data generation. Long et al. [14] review recent work in this area, organizing it into three main components: generation, curation, and evaluation. They propose a general workflow for LLM-driven synthetic data generation based on these components.

In the medical domain, NoteChat [5] introduces a novel cooperative multi-agent framework that utilizes Large Language Models (LLMs) to generate patient-physician dialogues. The framework consists of three modules: Planning, Roleplay, and Polish. It follows the principle that an ensemble of role-specific LLMs can more effectively fulfill their respective roles through structured role-play and strategic prompting.

doi:10.6342/NTU202504251

Dataset	# Dialogue	Avg. tokens per dialogue	Avg. utterances per dialogue
MedDialog-CN[3]	3,407,494.00	193.74	3.30
ReMeDi-large[4]	95,408.00	302.29	18.38
FamilyMed-Dialogue-Note [1]	94.00	5,779.84	284.78
ACI-Bench[2]	207.00	1,302.00	55.00

Table 3.1: Statistics comparison between datasets used and other related medical dialogue datasets.

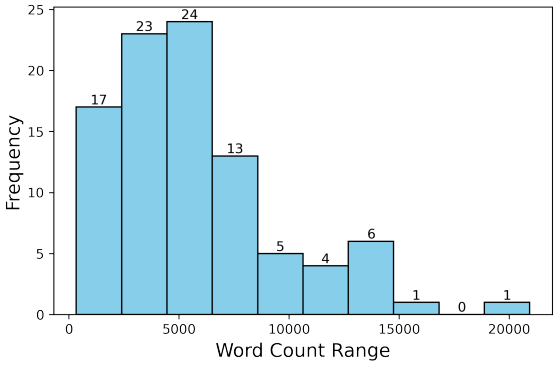


Figure 3.1: Distribution of Word Counts in Traditional Chinese FamilyMed-Dialogue-Note Dialogues

Chapter 3. Datasets

We adopt the FamilyMed-Dialogue-Note Dataset [1] and the ACI-Bench Dataset [2] for our experiments. Both contain outpatient doctor-patient conversations paired with corresponding clinical notes structured in the SOAP format (Subjective, Objective, Assessment, and Plan). Given our focus on real-world, privacy-sensitive applications in a local healthcare setting, we primarily emphasize the FamilyMed-Dialogue-Note Dataset, which was collected from National Taiwan University Hospital and better reflects the constraints and characteristics of our target use case.

doi:10.6342/NTU202504251

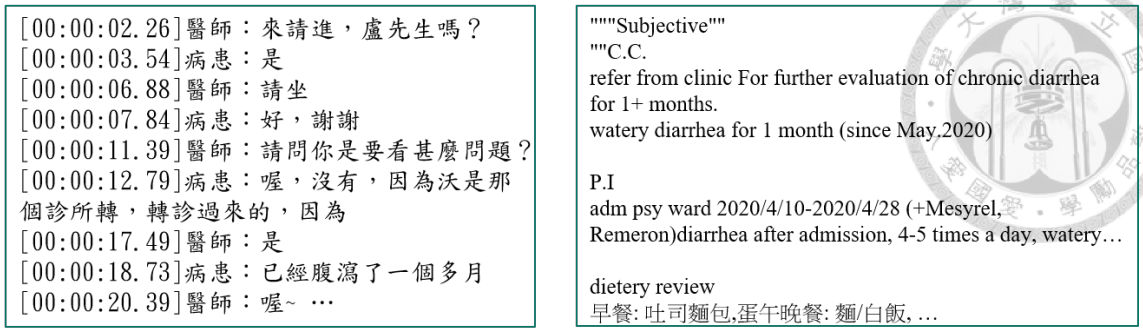


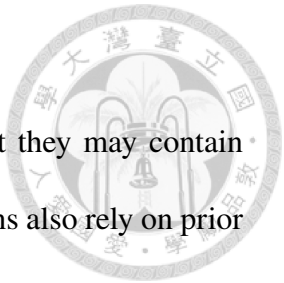
Figure 3.2: Part of an example of the FamilyMed-Dialogue-Note Dataset

3.1 FamilyMed-Dialogue-Note Dataset

The FamilyMed-Dialogue-Note dataset was collected by [1] in collaboration with teaching clinics of the Family Medicine Department at National Taiwan University Hospital (FamilyMed-Dialogue-Note). Each case includes a transcript of a real outpatient visit and an associated clinical note written in the SOAP format (Subjective, Objective, Assessment, and Plan). Figure 3.2 shows an example. The original dialogues are in Traditional Chinese, while the clinical notes are primarily written in English. Since other real-world data are not available, and publicly available synthetic medical dialogue datasets that are similar to our setting are mostly in English, we translate the FamilyMed-Dialogue-Note transcripts into English.

Unlike public medical dialogues collected from online platforms, the FamilyMed-Dialogue-Note dataset contains **longer**, more **colloquial** conversations that reflect actual outpatient visits in teaching clinics of the Family Medicine Department. Individual utterances are typically short, with meaningful information often spread across multiple turns. Figure 3.1 shows the #token distribution. However, a major challenge is the **limited scale** of the FamilyMed-Dialogue-Note dataset, which reflects the constraints of working with privacy-sensitive hospital data in a local setting. Table 3.1 presents comparative statistics between FamilyMed-Dialogue-Note, ACI-Bench, and other public

doi:10.6342/NTU202504251



medical dialogue datasets.

Another notable characteristic of real-world clinical notes is that they may contain information **not explicitly mentioned in the dialogue**, since physicians also rely on prior medical records, test results, or observations outside the conversation. Therefore, we re-annotated the dataset to identify and label information **not covered** in the dialogues. Moreover, the Objective part in the dataset is measurable data, such as vital signs, physical exam findings, and test results, which are rarely verbalized in dialogues. As a result, we only do experiments on the **Subjective**, **Assessment**, and **Plan** components, especially the Subjective section, which directly reflects the patient's spoken narrative. There are 14 structured columns from the clinical notes: *Chief complaint, History of present illness, Past medical history, Current medications, Allergy history, Family history, History of betel nut, History of drinking, Smoking history, Regular exercise, Profession, Diet, Assessment, and Plan.* Details regarding the translation, summarization, and re-annotation processes are described in Chapter 4.

3.2 ACI-Bench Dataset

The ACI-Bench dataset [2] was utilized in ACL ClinicalNLP MEDIQA-Chat 2023. It consists of three subsets, each representing a common mode of clinical note generation from doctor-patient conversations:

- **Virtual Assistant:** The doctor issues explicit commands to interact with a virtual assistant (e.g., "Hey Dragon, show me the diabetes labs") during the outpaient visit. This subset is created by a team of 5+ medical experts.

- **Virtual Scribe:** The doctor speaks naturally, possibly giving free-form instructions to a human or virtual scribe to assist in composing the clinical note.

doi:10.6342/NTU202504251

- **Ambient Clinical Intelligence (ACI):** A natural conversation takes place between a physician and a patient, without explicit commands to a virtual assistant or instructions directed at a scribe. This subset was created through role-playing between a certified physician and a layperson volunteer, based on a list of symptom prompts.

Clinical notes were initially generated by an automated note-generation system and subsequently reviewed and revised by domain experts. As shown in Table 3.1, among public medical dialogue datasets, ACI-Bench contains relatively longer conversations, making it more comparable to the FamilyMed-Dialogue-Note dataset. Additionally, the clinical notes are written in the SOAP format as well. Thus, we selected ACI-Bench to evaluate whether the experimental results are consistent with those obtained from the FamilyMed-Dialogue-Note dataset.

Chapter 4. Methodology

In this chapter, as shown in Figure 4.1 we introduce the data preprocessing steps for the FamilyMed-Dialogue-Note dataset and methods used for our main task, that is summarizing clinical notes from outpatient conversations. To ensure compatibility with privacy-sensitive hospital data in a fully local setting, all approaches are designed to work exclusively with open-source large language models (LLMs) that can be run on local machines. Moreover, evaluation metrics used are mentioned at the end of the section.

- **FamilyMed-Dialogue-Note Data preprocessing:** Includes details of dialogue translation, summarization and clinical notes' re-annotation.

- **Clinical Note Generation:** Includes One-stage End-to-end Generation, Two-stage

doi:10.6342/NTU202504251

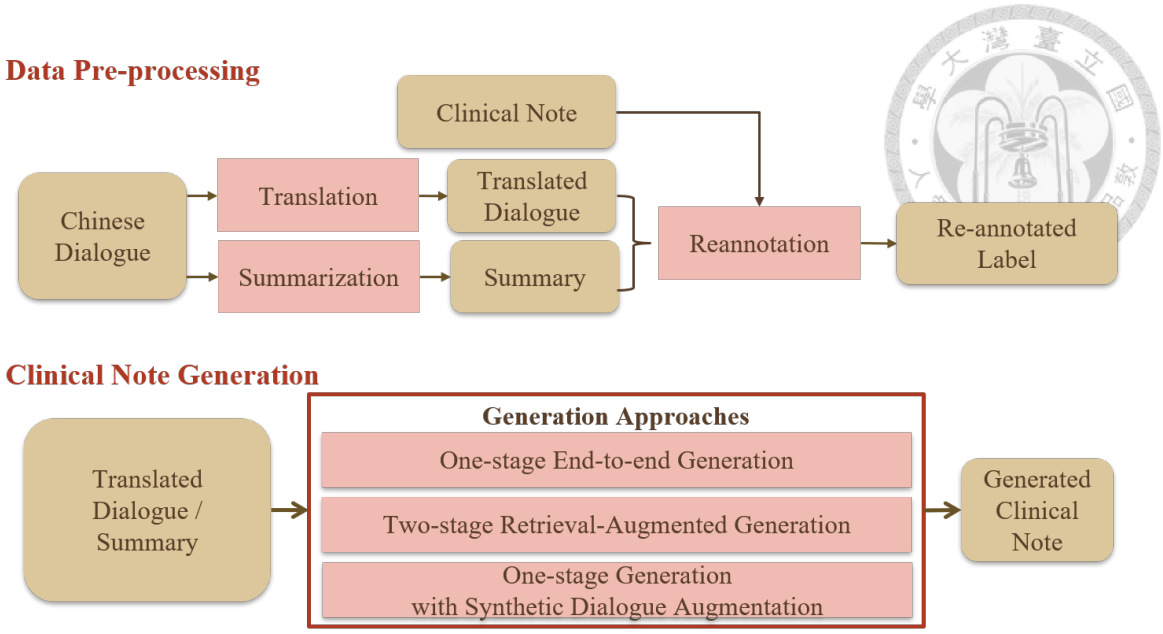


Figure 4.1: Overall workflow of the proposed methodology, from data preprocessing to the clinical note generation task.

Data	Min	Q1	Q2	Q3	Max	Mean
Orginal Chinese Dialogue	338.00	3250.50	5128.50	7301.75	20932.00	5779.84
Generated English Summary	96.00	608.50	994.50	1453.75	3911.00	1134.78
Translated English Dialogue	232.00	2110.25	3320.50	4551.50	13756.00	3747.33

Table 4.1: Token Count Statistics

Retrieval-Augmented Generation and One-stage Generation with Synthetic Dialogue Augmentation.

4.1 FamilyMed-Dialogue-Note Data Preprocessing

Initially, data cleaning was performed on both the conversation transcripts and the clinical notes. Time stamps were removed from the transcripts, and the unstructured notes were converted into structured data in JSON format. Subsequently, translation, summarization, and re-annotation were carried out as part of the data preparation process.

doi:10.6342/NTU202504251

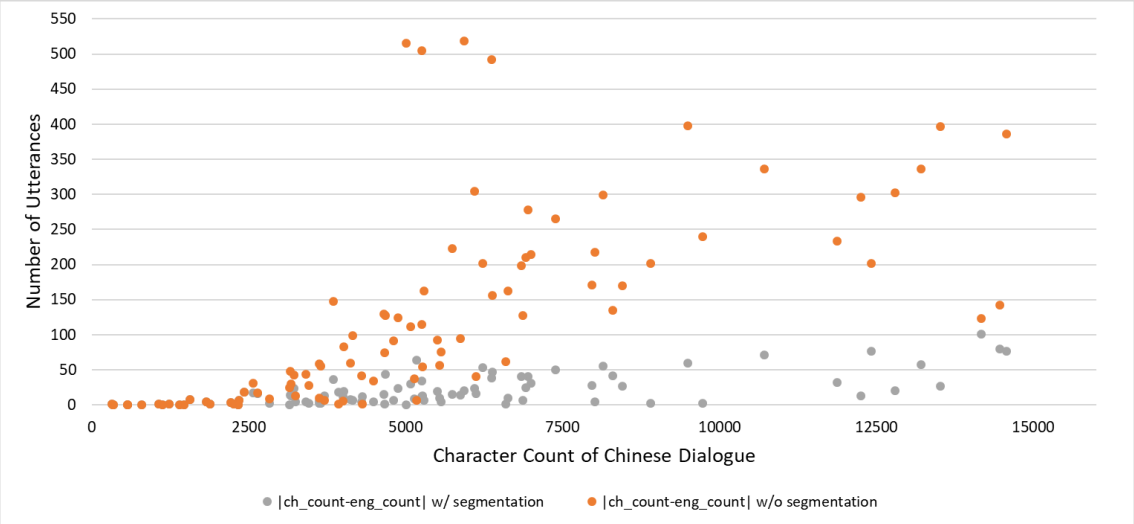


Figure 4.2: Absolute difference in dialogue turn counts between original and translated conversations, with and without segmentation. (Obvious outliers have been omitted.)

4.1.1 Translation

The original Chinese doctor-patient dialogues were translated into English using **LLaMA-3.3-70B-Instruct**. First, we input the entire conversation into the model for direct translation. However, as shown in Table 4.1, the dataset has an average of 5,779.84 tokens per dialogue, with 12 entries exceeding 10,000 Chinese characters. This led to the well-known "lost in the middle" issue, where the model struggles to maintain context in long sequences. For further investigation, we analyzed the number of utterances in each conversation before and after translation. The average difference in utterance count was 136.88, showing a substantial loss or hallucination of content. Orange dots in Figure 4.2 indicate that the difference is positively related to the conversation length.

To mitigate these issues, we then segmented the dialogues into chunks based on complete speaker turns, with each segment limited to approximately 2,500 characters. Also, A sliding window of five sentences was applied to maintain contextual continuity between segments. The translated segments were then concatenated to reconstruct the full English dialogue. Although the "lost in the middle" problem was not completely

doi:10.6342/NTU202504251

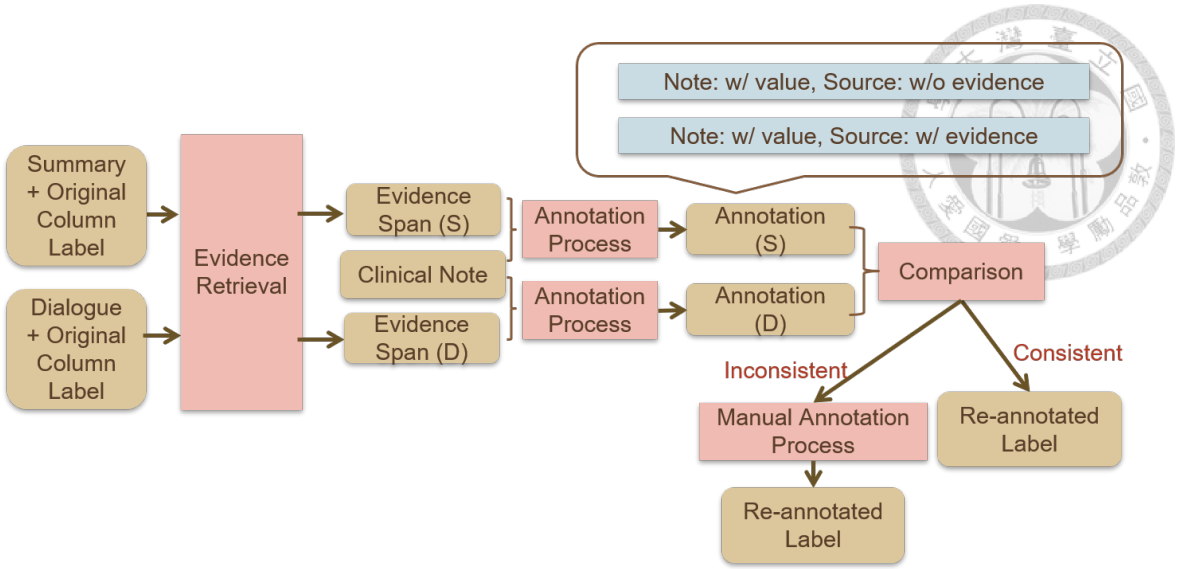


Figure 4.3: Workflow of the re-annotation process.

eliminated, the average difference in utterance count was significantly reduced to 19.97. Figure 4.2 clearly illustrates the improvement achieved through this segmentation strategy.

4.1.2 Summarization

We generate the English summaries directly from the original Chinese doctor-patient dialogues. Following the segmentation strategy described in Section 4.1.1, each dialogue was divided into segments of 1,000 characters with overlapping sliding windows to preserve context. Summaries generated by **LLaMA-3.3-70B-Instruct** and **LLaMA-3.1-8B-Instruct** exhibited a high similarity of 0.8970. The experiment results did not show big differences. In line with our low-resource, fully local setup, we selected the summaries generated by LLaMA-3.1-8B-Instruct for use in subsequent experiments. The data statistics are reported in Table 4.1.

doi:10.6342/NTU202504251

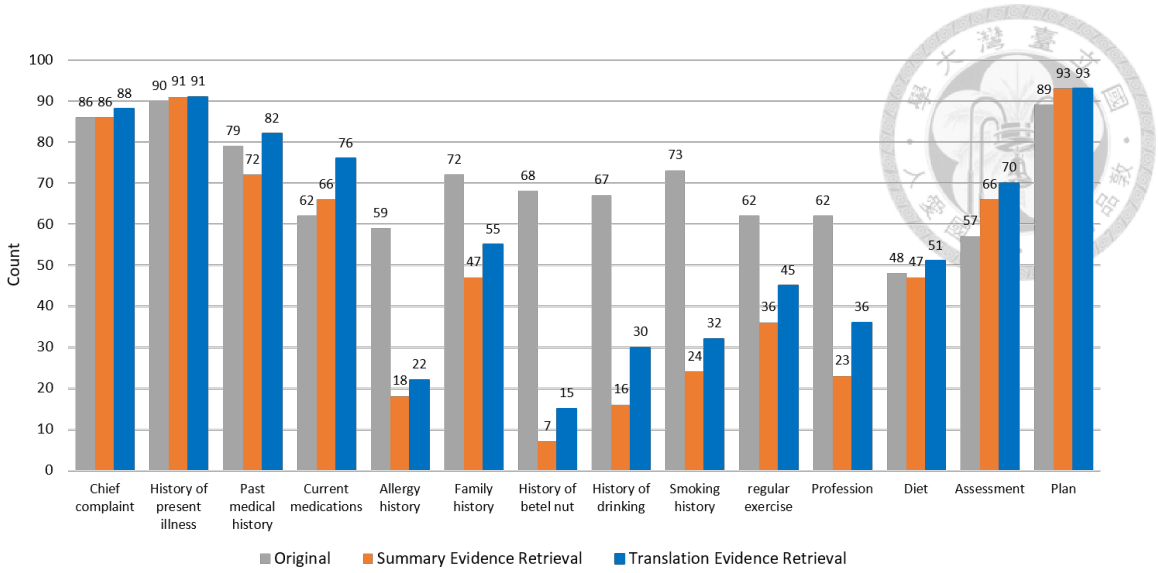
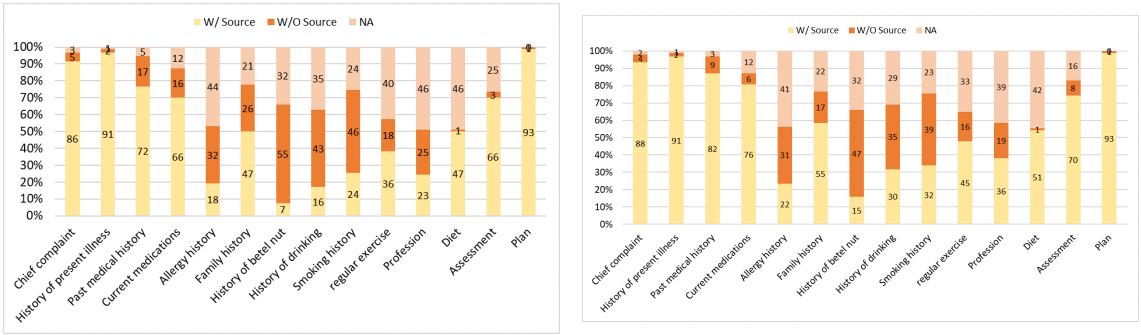


Figure 4.4: Number of entries with source information before and after evidence retrieval

4.1.3 Re-annotation



(a) Summary Evidence Retrieval

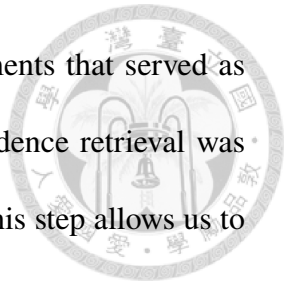
(b) Dialogue Evidence Retrieval

Figure 4.5: Percentage composition of annotations.

W/ Source: The note label's presence in original contexts is validated. W/O Source: The note label isn't found in original contexts. NA: The original label is 'na' and also no information found by retrieval.

As mentioned in 3.1, clinical notes in real-world outpatient settings often contain information that is not fully verbalized during the patient-doctor interaction. Therefore, we performed re-annotation to identify which parts of the clinical note were actually grounded in the dialogue and which were not. Re-annotation steps are shown in Figure 4.3. This process helps improve evaluation fairness and enables models to be trained or evaluated only on content that is truly present in the conversation.

doi:10.6342/NTU202504251



We used **LLaMA-3.3-70B-Instruct** to retrieve the specific segments that served as the evidence source for each labeled field in the clinical notes. Evidence retrieval was performed on both translation-based and summary-based datasets. This step allows us to trace the origin of note content.

As described in Section 4.1.1, segmentation was initially applied. To validate the presence of model-identified source sentences within their original contexts, we employed longest common subsequence matching, computing similarity scores via SequenceMatcher. The string matching similarity was found to be 0.4439 for the translation-based dataset and 0.8665 for summary-based datasets. The diminished score in the translation-based dataset did not indicate hallucinated content, but instead reflected challenges in string-level matching due to paraphrasing, punctuation alterations, or LLM-driven rewording. Manual verification confirmed the trustworthiness of these retrieved evidences, despite lexical variations. This process served to identify clearly absent sources.

Figure 4.4 illustrates the comparison of the number of data points with identified sources before and after the evidence retrieval process. The columns with the most consistently annotated sources are *Chief Complaint*, *History of Present Illness*, and *Plan*, while a substantial portion of entries for *Allergy History*, *History of Betel Nut*, *History of Drinking*, *Smoking History*, and *Profession* lacked identifiable label sources. To further improve annotation reliability, we compared evidence retrieval results between translated dialogues and their corresponding summaries, as shown in Figure 5a and Figure 5b. In cases where the model identified inconsistent sources between the two, we conducted manual re-annotation to confirm the correct label. This cross-validation step helped ensure high-quality annotations for downstream training and evaluation. Additionally, those columns lacked source references were also re-annotated by manual check.

doi:10.6342/NTU202504251

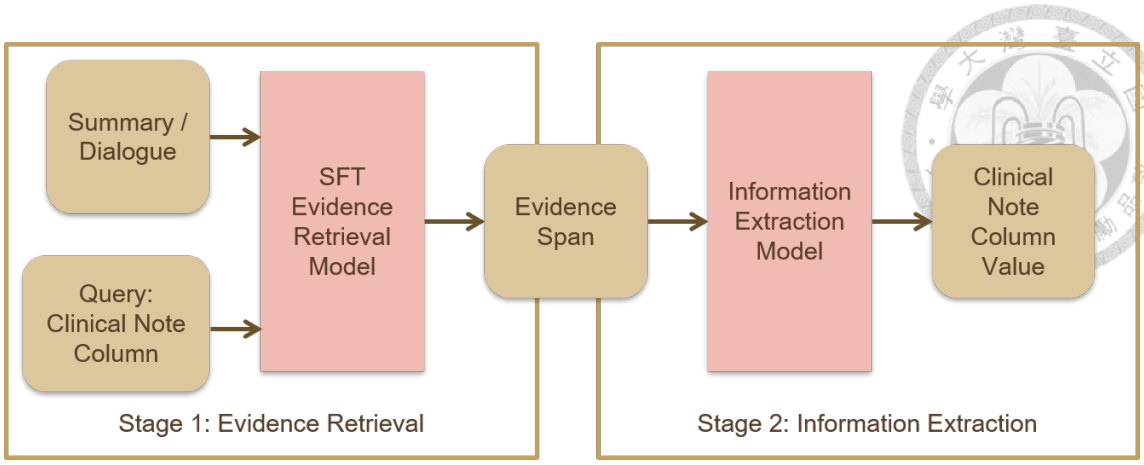


Figure 4.6: Overview of the Two-stage Retrieval-Augmented Generation

4.2 Clinical Note Generation

4.2.1 One-stage End-to-end Generation

For both the FamilyMed-Dialogue-Note and ACI-Bench datasets, we evaluated large language models (LLMs) under various configurations, including different model sizes, few-shot prompting, supervised fine-tuning, and supervised fine-tuning combined with few-shot prompting. To enable efficient fine-tuning with limited computational resources, we adopted Low-Rank Adaptation (LoRA). These experiments aimed to compare LLM performance across key prompting strategies and assess the impact of fine-tuning and instruction-following capabilities.

4.2.2 Two-stage Retrieval-Augmented Generation

This experiment was designed primarily to investigate whether supervised fine-tuning (SFT) enables LLMs to effectively capture and localize relevant information from doctor-patient conversations. While previous experiments focused on direct generation of clinical notes, the RAG framework provides a more interpretable two-stage pipeline that explicitly separates evidence retrieval from content generation. This design aims to prove whether

doi:10.6342/NTU202504251

fine-tuned models can reliably retrieve source spans that align with the clinical labels, based on the ground-truth established through re-annotation.

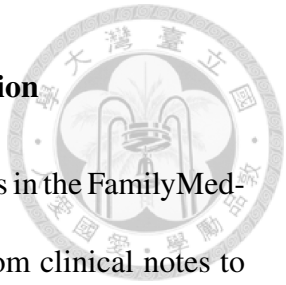
As in Figure 4.6, we constructed a two-stage Two-stage Retrieval-Augmented Generation framework using **LLaMA-3.1-8B-Instruct**. Both models in the pipeline were trained using supervision derived from the re-annotated FamilyMed-Dialogue-Note dataset.

Stage 1: Evidence Retrieval The first model in the RAG pipeline is trained to retrieve relevant evidence spans from the input dialogue or summary. Each training instance consists of a dialogue (or its summary) as input, and the corresponding evidence span(s) identified in the re-annotation stage as output. This model simulates a retrieval step by learning to locate the specific segments in the conversation that support each clinical note field.

Stage 2: Information Extraction The second model takes the retrieved evidence spans (i.e., the outputs of the first stage) as input and generates the final content for each structured field in the clinical note. During training, the input to this model is the retrieved evidence text, and the target output is the corresponding column value (e.g., 'Chief complaint', 'Family history', etc.) from the ground truth clinical note. This setup allows us to evaluate whether the model can extract accurate labels given only the relevant context.

By separating retrieval from generation, the RAG approach enhances both transparency and interpretability. Moreover, it functions as a diagnostic framework for assessing whether supervised fine-tuning (SFT) effectively enables the model to ground its outputs in the retrieved source context.

doi:10.6342/NTU202504251



4.2.3 One-stage Generation with Synthetic Dialogue Augmentation

To overcome the limited scale of real-world outpatient conversations in the FamilyMed-Dialogue-Note dataset, we explored synthetic dialogue generation from clinical notes to improve the diversity of training data. In particular, to address the imbalance caused by the high proportion of 'na' labels, we restricted generation to entries with "non-na" column labels.

We designed four generation strategies using **LLaMA-3.3-70B-Instruct** for better diversity:

1. Same Note (1-shot) Given a target clinical note and its corresponding dialogue, the goal is for the model to generate a synthetic dialogue that mimics the original's writing style.

2. Random Note-Dialogue (1-shot) To encourage diversity, we randomly sample a dialogue-note pair from the training set as an in-context example, regardless of its similarity to the target note. This approach introduces variation in prompting patterns and tests the model's generalization capability under mismatched examples.

3. Complex Guidance-Based Generation We designed a detailed set of instructions to guide the generation of highly realistic and nuanced conversations referring to the NoteChat framework [5]. The model is prompted with the clinical note and a comprehensive guideline that ensures the conversation follows natural clinical interaction flow. The guidance includes the following key elements:

- Start with a greeting from the doctor and maintain a turn-by-turn exchange.

doi:10.6342/NTU202504251

- Make the conversation colloquial, especially for the patient, while keeping the doctor's language more professional.
- Target 50—250 dialogue turns per conversation, often spreading a single piece of information over multiple turns.
- Ensure patient-reported symptoms align closely with the clinical note, and exclude information marked as "na".
- Ensure logical and progressive questioning by the doctor.
- Include small talk and realistic interjections to enhance authenticity.
- Conclude with a summary and plan from the doctor.

This method aims to generate richly detailed dialogues that more accurately reflect the complexity of real-world consultations.

4. Roleplaying-Based Generation To simulate natural and comprehensive doctor-patient interactions, the model is instructed to take on the roles of both doctor and patient. The generation prompt is initialized with role definitions and a clinical note, and the model then alternates between generating turns for each participant. Inspired by the NoteChat framework [5], we designed a mechanism to guide the conversation based on structured coverage. Specifically, we first convert the clinical note into a structured checklist of fields (e.g., chief complaint, family history). Initially, we attempted to enforce a fixed-format response to identify covered items by exact string comparison. However, this rigid strategy proved unreliable, as the model's responses often paraphrased or combined items, making it difficult to accurately detect coverage and causing conversations to either stall or continue indefinitely. As an alternative, at each turn, we apply a Coverage Checking

doi:10.6342/NTU202504251

Prompt to identify which items in the checklist have not yet been discussed. The output of the Coverage Checking Prompt was fed directly into the doctor and patient prompts as context. This allowed the model to organically incorporate the remaining topics into the ongoing conversation and to end more fluently and naturally.

4.3 Evaluation Metrics

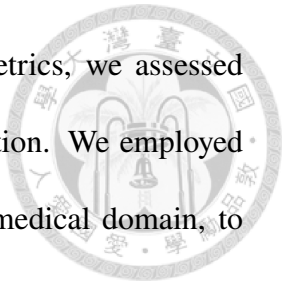
To evaluate the quality and clinical relevance of the generated notes, we employed a combination of automatic evaluation metrics and LLM-based semantic analysis. Our primary evaluation follows the guidelines and metrics used in the MEDIQA-Chat 2023 shared task [6].

Standard Generation Metrics. We adopted three widely used automatic metrics for evaluating text generation quality:

- **ROUGE-1:** Measures unigram (word-level) overlap between the generated note and the reference. It captures lexical similarity and is commonly used in summarization tasks.
- **BERTScore-F1:** Computes semantic similarity using contextual embeddings from a pre-trained BERT model. It better reflects meaning preservation, especially in medically rephrased content.
- **BLEURT:** A learned metric that combines fluency, grammar, and semantic accuracy, fine-tuned on human judgment data for sentence-level evaluation.

To provide an overall score, we also report the **average** of ROUGE-1, BERTScore-F1, and BLEURT, to balance lexical overlap, semantic similarity, fluency and correlation with human judgment.

doi:10.6342/NTU202504251



Medical Coverage and Observations. In addition to standard metrics, we assessed how well the generated notes covered the clinically relevant information. We employed **LLaMA3-OpenBioLLM-70B**, an advancing open-source LLM in medical domain, to provide structured explanations and identify:

- Whether source contexts are correctly preserved.
- Whether hallucinated or unsupported statements are present.

doi:10.6342/NTU202504251

Chapter 5. Experiments

5.1 Experimental Setup

In this thesis, we evaluate and compare the performance of these large language models (LLMs) of varying sizes and architectures: LLaMA-3.1-8B-Instruct, Mistral-Small-Instruct-2409, Gemma-2-27b-it and LLaMA-3.3-70B-Instruct. For clarity, we refer to these models as LLaMA-8B, Mistral-22B, Gemma-27B, and LLaMA-70B, respectively.

We conduct experiments on two datasets: the FamilyMed-Dialogue-Note dataset and the ACI-Bench dataset (see Section 3 for details). For the FamilyMed-Dialogue-Note dataset, we use 75 samples for training (with 10% held out for validation) and 19 samples for testing. The ACI-Bench dataset is split into 67 training samples, 20 validation samples, and 40 test samples.

To evaluate model performance, we adopt standard text generation metrics: ROUGE-1, BERTScore-F1, and BLEURT, and also report the average of these three scores , as described in Section 4.3. Further evaluation of clinical faithfulness and information grounding is conducted using LLaMA3-OpenBioLLM-70B explanations, which are discussed in Section 6.

All experiments are performed with at most 2 48G NVIDIA RTX A6000 GPUs, depending on model size and memory requirements. Under limited GPU memory constraints, supervised fine-tuning is performed on LLaMA-8B using Low-Rank Adaptation (LoRA), which enables efficient parameter-efficient training.

We organize our experiments around three clinical note generation strategies described

doi:10.6342/NTU202504251

LLM	ROUGE-1	BERTScore-F1	BLEURT	Average
LLaMA-8B	0.4855	0.7022	0.3750	0.5209
Mistral-22B	**0.5197**	**0.7195**	**0.4039**	**0.5477**
Gemma-27B	0.5059	0.7065	0.3460	0.5195
LLaMA-70B	0.4840	0.7013	0.3787	0.5213

Table 5.1: Evaluation of LLMs of different sizes (7B to 70B) on clinical dialogue generation (FamilyMed-Dialogue-Note).

LLM	ROUGE-1	BERTScore-F1	BLEURT	Average
LLaMA-8B	0.2188	0.5235	**0.4937**	0.4120
Mistral-22B	0.2815	0.5553	0.4027	0.4132
Gemma-27B	0.3723	**0.6210**	0.4390	0.4774
LLaMA-70B	**0.3810**	0.6132	0.4572	**0.4838**

Table 5.2: Evaluation of LLMs of different sizes (7B to 70B) on clinical dialogue generation (ACI-Bench).

in Section 4.2:

- **One-stage End-to-end Generation:** Leveraging LLMs to compare zero-shot inference across different model sizes, few-shot prompting strategies, and supervised fine-tuning applied to both summaries and dialogues.
- **Two-stage Retrieval-Augmented Generation:** Utilizing retrieved relevant segments from the dialogue / summary to generate specific clinical note fields.
- **One-stage Generation with Synthetic Dialogue Augmentation:** Generating artificial patient-doctor conversations from clinical notes to augment training data.

doi:10.6342/NTU202504251

5.2 One-stage End-to-end Generation

5.2.1 Impact of Model Size

We first examine the impact of model size on the clinical note generation task. Table 5.1 presents the performance of four LLMs of varying sizes, ranging from 8B to 70B parameters, on the FamilyMed-Dialogue-Note dataset. Among the evaluated models, Mistral-22B achieves the highest performance across all metrics, with an average score of 0.5477, indicating strong alignment with reference clinical notes in both lexical overlap and semantic similarity. Interestingly, LLaMA-70B, despite being the largest model, does not outperform its smaller counterparts, suggesting that model size alone does not guarantee better generation quality in this low-resource, domain-specific setting.

In contrast, results on the ACI-Bench dataset, shown in Table 5.2, reveal a clearer benefit from larger model sizes. LLaMA-70B attains the highest overall average score, driven by strong ROUGE-1 and BLEURT performance, while Gemma-27B achieves the highest BERTScore-F1, reflecting strong semantic similarity to the reference notes. Smaller models like LLaMA-8B and Mistral-22B trail behind, with particularly low ROUGE-1 scores, suggesting reduced ability to capture surface-level content overlap.

These findings highlight that model size alone is not the sole determinant of performance. In low-resource scenarios, moderately sized models like Mistral-22B can strike a favorable balance between performance and computational cost, while larger models may offer additional benefits for datasets with richer or more diverse clinical content.

doi:10.6342/NTU202504251

Method	ROUGE-1	BERTScore-F1	BLEURT	Average
0-shot	**0.5197**	**0.7195**	**0.4039**	**0.5477**
3-shot	0.4691	0.6824	0.3916	0.5144

Table 5.3: Performance comparison of Mistral-22B under 0-shot and 3-shot settings (FamilyMed-Dialogue-Note).

Method	ROUGE-1	BERTScore-F1	BLEURT	Average
0-shot	0.3810	0.6132	0.4572	0.4838
2-shot	**0.3920**	**0.6382**	**0.4777**	**0.5026**

Table 5.4: Performance comparison of LLaMA-70B under 0-shot and 2-shot settings (ACI-Bench).

5.2.2 Effect of Few-Shot Prompting

To evaluate the adaptability of LLMs in few-shot scenarios, we conducted additional experiments with providing a small number of in-context examples during inference. These experiments were conducted on the best-performing models from the model size comparison: Mistral-22B on the FamilyMed-Dialogue-Note dataset and LLaMA-70B on the ACI-Bench dataset.

As shown in Table 5.3, Mistral-22B performed best in the 0-shot setting, achieving an average score of 0.5477, outperforming the 3-shot setup, which reaches 0.5144. This unexpected performance drop may be attributed to the increased input length and contextual noise introduced by additional examples, which could confuse the model in a low-resource domain with constrained lexical patterns. In contrast, Table 5.4 demonstrates that LLaMA-70B benefits from few-shot prompting on the ACI-Bench dataset. The 2-shot configuration yields an improved average score of 0.5026, compared to 0.4838 in the 0-shot setting. Notably, the largest gains appear in BERTScore-F1 and BLEURT, indicating that additional examples help the model better capture semantic alignment and

doi:10.6342/NTU202504251

Method	ROUGE-1	BERTScore-F1	BLEURT	Average
0-shot	0.4855	0.7022	0.3750	0.5209
SFT	**0.6128**	**0.7626**	0.3892	**0.5882**
Mistral-22B 0-shot	0.5197	0.7195	**0.4036**	0.5476

Table 5.5: Performance comparison highlighting the effect of supervised fine-tuning (SFT) on LLaMA-8B (FamilyMed-Dialogue-Note).

Method	ROUGE-1	BERTScore-F1	BLEURT	Average
0-shot	0.2188	0.5235	0.4937	0.4120
SFT	**0.5674**	**0.7514**	**0.5484**	**0.6224**
LLaMA-70B 2-shot	0.3920	0.6382	0.4777	0.5026

Table 5.6: Performance comparison highlighting the effect of supervised fine-tuning (SFT) on LLaMA-8B (ACI-Bench).

naturalness in note generation.

Overall, the effectiveness of few-shot prompting appears to be highly dependent on both the dataset and the model. While few-shot examples can enhance performance in some contexts, they may also hinder generation in others, especially when the model is already well-aligned with the task in zero-shot settings or when the dataset's structure is sensitive to input length and format.

5.2.3 Supervised Fine-tuning (SFT)

To further evaluate the effectiveness of supervised fine-tuning (SFT) for clinical note generation, we fine-tuned LLaMA-8B using Low-Rank Adaptation (LoRA) on both the FamilyMed-Dialogue-Note and ACI-Bench datasets.

According to Table 5.5, for the FamilyMed-Dialogue-Note dataset, **LLaMA-8B with SFT** achieved the best overall performance among all settings, outperforming both its 0-shot baseline and the best 0-shot result from Mistral-22B. Specifically, SFT improved ROUGE-1 from 0.4855 to 0.6128, BERTScore-F1 from 0.7022 to 0.7626, and the average

doi:10.6342/NTU202504251

Input Set	ROUGE-1	BERTScore-F1	BLEURT	Average
Summary	**0.6128**	**0.7626**	**0.3892**	**0.5882**
Dialogue	0.6064	0.7459	0.3590	0.5704

Table 5.7: Performance comparison of LLaMA-8B models trained on summary vs. dialogue inputs (FamilyMed-Dialogue-Note).

score from 0.5209 to 0.5882. While Mistral-22B in 0-shot setting achieved a higher BLEURT score, the overall gain in the other metrics clearly highlights the benefit of task-specific fine-tuning.

The improvements are even more pronounced on the ACI-Bench dataset (Table 5.6), where SFT boosted the average score from 0.4120 (0-shot) to 0.6224. LLaMA-8B with SFT also surpassed the best few-shot result from LLaMA-70B, which achieved an average score of 0.5026. These results demonstrate that supervised fine-tuning with LoRA can enable smaller models to outperform larger ones under prompting-only settings. All results highlights the strong highlighting the effectiveness of SFT on the clinical note generation task.

Overall, the results underscore the effectiveness of SFT in enhancing model performance and domain adaptation, even when applied to relatively compact models like LLaMA-8B.

5.2.4 Input Format Comparison

All results presented above are derived from the summary input set. To further assess the impact of input format, we evaluated the SFT performance using both summary and dialogue input sets. Table 5.7 illustrates that the model trained on summary inputs consistently outperforms the model trained on dialogue inputs across all evaluation metrics. These findings suggest that concise and structured input offers more effective supervision

doi:10.6342/NTU202504251

Method	ROUGE-1	BERTScore-F1	BLEURT	Average
SFT	**0.6128**	**0.7626**	**0.3892**	**0.5882**
SFT w/ 3-shot	0.5949	0.7458	0.3651	0.5686

Table 5.8: Performance comparison of LLaMA-8B models fine-tuned with or without few-shot prompting (FamilyMed-Dialogue-Note).

Method	ROUGE-1	BERTScore-F1	BLEURT	Average
SFT	**0.5674**	**0.7514**	**0.5484**	**0.6224**
SFT w/ 3-shot	0.5262	0.7327	0.5175	0.5921

Table 5.9: Performance comparison of LLaMA-8B models fine-tuned with or without few-shot prompting (ACI-Bench).

for clinical note generation.

5.2.5 Combining SFT with Few-Shot Prompts

Moreover, we explored the effect of few-shot prompting in the SFT setting using in-context examples. However, few-shot prompting did not provide additional benefit over direct SFT, and in fact resulted in slightly degraded performance (see Table 5.8 and Table 5.9). These results suggest that incorporating few-shot examples during fine-tuning may introduce unnecessary complexity or noise, potentially leading to marginal reductions in performance.

Information Source	rouge1	BERTScore-F1	BLEURT	Average
FamilyMed-Dialogue-Note	0.7676	0.8490	0.7929	0.7577
ACI-Bench	0.4926	0.6605	0.5120	0.5550

Table 5.10: Evidence retrieval model (SFT-LLaMA-8B) performance

doi:10.6342/NTU202504251

Retrieved Evidence Span

A patient, Mr. Lu, visits a doctor's office with a complaint **of abdominal pain and diarrhea** that has persisted for **over a month**. He has continued to experience diarrhea and abdominal distension since his discharge. He has also noticed that his abdominal pain has become more severe and his stool has become more **watery**. …

Column Label

Refer from clinic for further evaluation of **chronic diarrhea** for **1+ months**.
watery diarrhea for 1 month (since May.2020)

Figure 5.1: Example of a "Chief Complaint" retrieved evidence span.

Source Set	Stage-2 Information Extraction Method	ROUGE-1	BERTScore-F1	BLEURT	Average
Summary	Full-3-shot_70B	**0.5220**	**0.7137**	0.4051	**0.5469**
	Note-3-shot_70B	0.4956	0.6942	**0.4065**	0.5321
	SFT	0.4701	0.6795	0.3345	0.4947
	SFT(note-3-shot)	0.3852	0.6300	0.2777	0.4309
Dialogue	SFT	0.4239	0.6551	0.3104	0.4631
	SFT(note-3-shot)	0.5100	0.6979	0.3175	0.5085

Table 5.11: Performance comparison of Stage-2 extraction methods with summary and dialogue inputs (FamilyMed-Dialogue-Note).

5.3 Two-stage Retrieval-Augmented Generation

The RAG pipeline employed in this study consists of two stages: (1) an **evidence retrieval** stage, followed by (2) an **information extraction** stage. A shared retriever model is used across all RAG methods, and its retrieval quality is summarized in Table 5.10. An illustrative example of the retrieved evidence spans is provided in Figure 5.1. In the information extraction stage, we evaluate inference using LLaMA-70B with full few-shot (dialogue-note pairs) or note-only few-shot, as well as fine-tuning LLaMA-8B with or without additional few-shot prompts.

On the FamilyMed-Dialogue-Note dataset (Table 5.11), the best average performance (0.5469) is achieved by LLaMA-70B using full few-shot prompting with summary-based inputs. This method also leading in ROUGE-1 (0.5220) and BERTScore-F1 (0.7137), indicating strong lexical and semantic alignment. Overall, summary source sets outperform

doi:10.6342/NTU202504251

Method	ROUGE-1	BERTScore-F1	BLEURT	Average
SFT(Evidence Retrieval) + Full-3-shot_70B	0.5220	0.7137	**0.4051**	0.5469
One-Stage SFT	**0.6128**	**0.7626**	0.3892	**0.5882**

Table 5.12: Comparison of the best RAG-based method and the best one-stage SFT model (FamilyMed-Dialogue-Note).

Stage-2 Information Extraction Method	ROUGE-1	BERTScore-F1	BLEURT	Average
Full-3-shot_70B	0.3719	0.6289	0.4460	0.4823
Note-3-shot_70B	0.3728	0.6283	0.4548	0.4853
SFT	0.4026	0.6710	0.4600	0.5112
SFT(note-3-shot)	**0.4034**	**0.6816**	**0.4678**	**0.5176**

Table 5.13: Performance comparison of Stage-2 extraction methods (ACI-Bench).

dialogue-based ones. However, the SFT-LLaMA-8B model with note-3-shot prompting performs well on dialogue-based generation, achieving competitive ROUGE-1 and BERTScore-F1 compared to top summary-based methods.

When comparing RAG with the best one-stage SFT method (Table 5.12), we observe that the best two-stage RAG method (SFT(Evidence Retrieval) + full-few-shot with LLMA-70B) performs slightly worse overall. While RAG leads in BLEURT (0.4051 vs. 0.3892), indicating stronger semantic fluency, SFT consistently achieves higher ROUGE-1 and BERTScore-F1, suggesting better lexical and structural alignment.

For the ACI-Bench dataset, the best performance across all metrics is achieved by the SFT-LLaMA-8B model using note-3-shot prompting, unlike the FamilyMed-Dialogue-Note dataset where full-few-shot prompting with 70B performed best. Details are presented in Table 5.13.

On the whole, although the best two-stage RAG approach performs slightly worse than the one-stage SFT model, the results demonstrate the capability of smaller fine-tuned LLMs to retrieve and utilize evidence effectively. This suggests that when a model performs well in a one-stage setting, introducing a two-stage pipeline may lead to additional error

doi:10.6342/NTU202504251

Method	Self-BLEU(↓)	Factuality
Same Note (1-shot)	**0.7034**	0.5352
Random Note-Dialogue (1-shot)	0.7273	**0.8847**
Complex Guidance-Based Generation	0.7344	0.6580
Roleplaying-Based Generation	0.7510	0.4349

Table 5.14: Evaluation of generated dialogues (FamilyMed-Dialogue-Note).

propagation that offsets potential gains.

5.4 One-stage Generation with Synthetic Dialogue Augmentation

To enhance training data, we generated synthetic dialogues and filtered high-quality examples using Self-BLEU (for diversity) and factuality scores provided by LLaMA3-OpenBioLLM-70B. The average evaluation results of four generation methods are presented in Table 5.14. It is challenging to reflect the complexities of real-world conversations by synthetic dialogues from local LLMs:

- Same Note-Dialogue (1-shot), Random Note-Dialogue (1-shot): Generated sentences often lack diversity in structure and phrasing.
- Complex Guidance-Based Generation: Produces more complex outputs but does not follow all the guidelines.
- Roleplaying-Based Generation: Struggles to end conversations naturally.

Random Note-Dialogue (1-shot) and **Complex Guidance-Based Generation** had better balance in the scores, and samples have less hallucination explained by LLaMA3-OpenBioLLM-70B, so we selected the generated data by these 2 methods for augmentation.

Additionally, summaries corresponding to the selected synthetic dialogues were generated and included in training. As shown in Table 5.15, **SFT(Original + Complex**

doi:10.6342/NTU202504251

Generation Method	ROUGE-1	BERTScore-F1	BLEURT	Average
SFT(Original + Random Shot Generation)	0.5511	0.7179	0.3548	0.5413
SFT(Original + Complex Guidance Generation)	0.5840	0.7472	0.3843	0.5718
SFT(Original + All Generation)	0.3750	0.6309	0.3082	0.4381
SFT(Original)	**0.6128**	**0.7626**	**0.3892**	**0.5882**

Table 5.15: Performance of SFT models trained with synthetic data generated by different augmentation methods, compared to the baseline using original data only (FamilyMed-Dialogue-Note).

Guidance Generation) achieved the best performance among all augmentation strategies. However, overall metric scores did not improve compared to fine-tuning solely on original data. Furthermore, we observed a significant performance drop in columns with a high proportion of 'na' labels. This finding suggests that while the model's tendency to overgenerate 'na' labels for these columns may have changed, its ability to judge information coverage diminished.

doi:10.6342/NTU202504251

Chapter 6. Discussion

6.1 Case Study

To further interpret and evaluate the quality of generated clinical note fields, we conducted a case study based on outputs from our primary approaches. These analyses leveraged LLaMA3-OpenBioLLM-70B to provide explanations and comparative insights. Representative examples are presented in Table 6.1. Methods like LLaMA-8B - SFT (M1), Mistral-22B - inference (M2), LLaMA-70B - inference (M4), and the RAG method successfully captured essential information details. For instance, in the Chief Complaint example, they correctly identified ”1 month” duration, and in the Family History example, they accurately reported family members with corresponding conditions. The answers generated by LLaMA-8B - SFT exhibit the highest stylistic similarity to the corresponding column labels. The RAG method, in particular, demonstrated its efficacy by retrieving the correct evidence spans. In contrast, the SFT model with data augmentation (AUG) introduced inaccuracies, such as ”3 day” instead of ”1 month” in the Chief Complaint and incorrect family members with hallucinated histories in the Family History example. Note that while the ”niece” detail in the Family History for some models (M3, M4, RAG) is not in the label, it is present in the original dialogue, indicating faithful extraction rather than hallucination.

6.2 Column-Wise Analysis

As detailed in Table 6.2, a column-wise analysis of the three leading methods reveals distinct performance strengths and weaknesses across different parts of the clinical note. The results indicate that no single method is universally superior, and the best-performing

doi:10.6342/NTU202504251

Column	Method	Genetated Answer
Chief complaint	M1	abdominal pain and diarrhea for 1 month
	M2	Abdominal pain and diarrhea for over a month
	M3	Abdominal discomfort and constipation
	M4	Abdominal pain and diarrhea for over a month
	RAG	Abdominal pain and diarrhea for over a month, with increased severity and watery stools, accompanied by bloating and abdominal discomfort.
	AUG	Abdominal pain and constipation for 3 day
	Label	**Refer from clinic for further evaluation of chronic diarrhea for 1+ months. watery diarrhea for 1 month (since May.2020)**
Family history	M1	grandfather: liver cancer, grandmother: DM, stroke
	M2	Grandfather had liver cancer, grandmother had diabetes, a stroke, niece diagnosed with a neurological disorder at a young age
	M3	na
	M4	Grandfather had liver cancer, grandmother had diabetes, a stroke, and niece was diagnosed with a neurological disorder
	RAG	Neurological disorder (niece), liver cancer (grandfather), diabetes (grandmother), stroke (grandmother)
	AUG	grandfather: colon cancer, grandfather: stroke, grandmother: DM, father: HTN, mother: HTN
	Label	**grandfather: liver cancer, grandmother: DM, died of stroke**

Table 6.1: Case study of generated answers from main approaches.

M1: LLaMA-8B - SFT, M2: Mistral-22B - inference, M3: Gemma-27B - inference, M4: LLaMA-70B - inference, RAG: SFT(Evidence Retrieval) + Full-3-shot 70B, AUG: SFT(Original + Complex Guidance Generation).

doi:10.6342/NTU202504251

Column	Method	ROUGE-1	BERTScore-F1	BLEURT	Column Average
Chief complaint	One-stage SFT	0.2177	0.5381	0.2548	0.3369
	Evidence Retrieval + Full-3-shot_70B Extraction	0.1786	0.5223	0.3036	0.3348
	SFT(Original + Complex Guidance Generation)	**0.2592**	**0.5729**	**0.3274**	**0.3865**
History of present illness	One-stage SFT	0.1425	0.4839	0.2495	0.2920
	Evidence Retrieval + Full-3-shot_70B Extraction	0.1566	0.5198	**0.3563**	**0.3442**
	SFT(Original + Complex Guidance Generation)	**0.1746**	**0.5228**	0.2995	0.3323
Past medical history	One-stage SFT	0.3741	0.6320	0.2507	0.4189
	Evidence Retrieval + Full-3-shot_70B Extraction	0.4078	**0.6703**	**0.3194**	**0.4658**
	SFT(Original + Complex Guidance Generation)	**0.4211**	0.6285	0.2415	0.4304
Current medications	One-stage SFT	**0.5789**	**0.7236**	0.3430	0.5485
	Evidence Retrieval + Full-3-shot_70B Extraction	0.2632	0.5371	0.2142	0.3382
	SFT(Original + Complex Guidance Generation)	0.5789	0.7218	**0.3450**	**0.5486**
Allergy history	One-stage SFT	0.8571	0.9000	0.4752	0.7441
	Evidence Retrieval + Full-3-shot_70B Extraction	0.7694	0.8305	**0.4983**	0.6994
	SFT(Original + Complex Guidance Generation)	**0.8632**	**0.9053**	0.4796	**0.7494**
Family history	One-stage SFT	0.7368	0.8065	0.4626	0.6686
	Evidence Retrieval + Full-3-shot_70B Extraction	**0.7871**	**0.8596**	**0.5234**	**0.7234**
	SFT(Original + Complex Guidance Generation)	0.6564	0.7705	0.3739	0.6003
History of betel nut, drinking and smoking	One-stage SFT	**0.9474**	**0.9642**	0.5206	**0.8107**
	Evidence Retrieval + Full-3-shot_70B Extraction	0.8969	0.9395	**0.5233**	0.7866
	SFT(Original + Complex Guidance Generation)	0.8618	0.9212	0.4961	0.7597
Regular exercise	One-stage SFT	**0.6488**	**0.7847**	0.3822	0.6052
	Evidence Retrieval + Full-3-shot_70B Extraction	0.6269	0.7734	**0.4377**	**0.6127**
	SFT(Original + Complex Guidance Generation)	0.6182	0.7713	0.3983	0.5959
Profession	One-stage SFT	0.7424	0.8367	0.4529	0.6773
	Evidence Retrieval + Full-3-shot_70B Extraction	0.7382	0.8449	0.4493	0.6775
	SFT(Original + Complex Guidance Generation)	**0.7579**	**0.8384**	**0.4577**	**0.6847**
Diet	One-stage SFT	**0.6842**	**0.7970**	0.4016	**0.6276**
	Evidence Retrieval + Full-3-shot_70B Extraction	0.5569	0.7280	**0.4148**	0.5666
	SFT(Original + Complex Guidance Generation)	0.5711	0.7432	0.3888	0.5677
Assessment	One-stage SFT	**0.5789**	**0.7243**	**0.3326**	**0.5453**
	Evidence Retrieval + Full-3-shot_70B Extraction	0.0327	0.3793	0.2790	0.2303
	SFT(Original + Complex Guidance Generation)	0.5263	0.6868	0.3085	0.5072
Plan	One-stage SFT	**0.1752**	**0.5571**	0.2820	**0.3381**
	Evidence Retrieval + Full-3-shot_70B Extraction	0.0993	0.5078	**0.3046**	0.3039
	SFT(Original + Complex Guidance Generation)	0.1641	0.5361	0.2713	0.3238

Table 6.2: Comparison of the best-performing methods from each of the three approaches (One-stage End-to-end Generation, Two-stage Retrieval-Augmented Generation , and One-stage Generation with Synthetic Dialogue Augmentation) across all clinical note columns.

doi:10.6342/NTU202504251

approach often depends on the specific type of information being generated. A key finding is the overall **stability of the One-stage SFT method**, which provides a consistent performance baseline across all columns.

6.2.1 Chief Complaint

For the **Chief Complaint** column, the **SFT(Original + Complex Guidance Generation)** method achieved the highest scores across all metrics, likely benefiting from the diverse phrasings introduced during the augmentation process.

6.2.2 Patient History and Lifestyle

In the various patient history columns, performance was more distributed. The **Evidence Retrieval + Full-3-shot_70B Extraction** method was the strongest for generating the **History of present illness, Past medical history and Family history** particularly excelling on BERTScore-F1 and BLEURT. This indicates that grounding the model with retrieved evidence is highly effective for synthesizing factual, context-dependent information. Conversely, for more straightforward and list-based lifestyle-related clinical note entries, the simpler One-stage SFT approach demonstrated the highest performance. This suggests that for columns such as **History of betel nut, drinking and smoking, Regular exercise and Diet**, a direct fine-tuning approach without augmentation or retrieval is often sufficient and less prone to introducing errors. The simplicity of these labels likely contributed to the particularly high scores achieved on these columns.

6.2.3 Assessment and Plan

The most significant performance differences were observed in the **Assessment** and **Plan** columns, which require a high degree of clinical judgment and synthesis. The

doi:10.6342/NTU202504251

One-stage SFT method performed exceptionally well, achieving the highest scores for both **Assessment** (Average: 0.5453) and **Plan** (Average: 0.3381). Notably, the **Evidence Retrieval + Full-3-shot_70B Extraction** method performed very poorly on both columns. This highlights a critical limitation of the two-stage RAG approach, where error propagation from the initial retrieval phase may negatively impact the final output, particularly for tasks that require complex inference rather than straightforward information extraction.

doi:10.6342/NTU202504251

Chapter 7. Conclusion

This work focused on the local application of Large Language Models for clinical note generation. Our methodology involved a robust data preprocessing pipeline, including summarization, translation, and re-annotation. We further investigated three distinct note generation paradigms: One-stage End-to-end Generation, Two-stage Retrieval-Augmented Generation (RAG), and One-stage Generation with One-stage Generation with Synthetic Dialogue Augmentation.

Our experiments demonstrated the effectiveness of supervised fine-tuning (SFT) methods and the capability of smaller LLMs in retrieving accurate evidence spans. Notably, the **one-stage SFT approach offers a strong balance between performance and computational efficiency**, achieving stable results while requiring significantly lower inference cost compared to multi-stage or large-model few-shot pipelines.

On the other hand, synthetic data proved unstable in faithfully reflecting the complexities of real-world conversations. The complexity and length of real-world clinical contexts also contributed to performance limitations across all methods.

Overall, our findings highlight that cost-effective, locally deployable LLMs, especially when well fine-tuned, can effectively assist in summarizing clinical notes from conversations. However, challenges remain, including long-context handling, data imbalance, evaluation limitations, and the constraints of fully local deployment. Future work should focus on scaling datasets, enhancing faithfulness evaluation, and involving domain experts to strengthen clinical applicability.

doi:10.6342/NTU202504251

References

[1] Hsin-Yu Tsai, Hen-Hsen Huang, Che-Jui Chang, Jaw-Shiun Tsai, and Hsin-Hsi Chen. Patient history summarization on outpatient conversation. In 2022 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), pages 364–370. IEEE, 2022.

[2] Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. Scientific data, 10(1):586, 2023.

[3] Xuehai He, Shu Chen, Zeqian Ju, Xiangyu Dong, Hongchao Fang, Sicheng Wang, Yue Yang, Jiaqi Zeng, Ruisi Zhang, Ruoyu Zhang, et al. Meddialog: Two large-scale medical dialogue datasets. arXiv preprint arXiv:2004.03329, 2020.

[4] Guojun Yan, Jiahuan Pei, Pengjie Ren, Zhaochun Ren, Xin Xin, Huasheng Liang, Maarten De Rijke, and Zhumin Chen. Remedi: Resources for multi-domain, multi-service, medical dialogues. In Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 3013–3024, 2022.

[5] Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. Notechat: a dataset of synthetic doctor-patient conversations conditioned on clinical notes. arXiv preprint arXiv:2310.15959, 2023.

[6] Asma Ben Abacha, Wen-wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen-Yildiz. Overview of the mediqa-chat 2023 shared tasks on the summarization & generation of doctor-patient conversations. In Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 503–513, 2023.

[7] Marcus V Ortega, Michael K Hidrue, Sara R Lehrhoff, Dan B Ellis, Rachel C Sisodia, William T Curry, Marcela G Del Carmen, and Jason H Wasfy. Patterns in physician

doi:10.6342/NTU202504251

burnout in a stable-linked cohort. JAMA Network Open, 6(10):e2336745–e2336745, 2023.

[8] Jeffrey Budd. Burnout related to electronic health record use in primary care. Journal of primary care & community health, 14:21501319231166921, 2023.

[9] Yizhan Li, Sifan Wu, Christopher Smith, Thomas Lo, and Bang Liu. Improving clinical note generation from complex doctor-patient conversation. In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pages 209–221. Springer, 2025.

[10] John Giorgi, Augustin Toma, Ronald Xie, Sondra S Chen, Kevin R An, Grace X Zheng, and Bo Wang. Wanglab at mediqa-chat 2023: Clinical note generation from doctor-patient conversations using large language models. arXiv preprint arXiv:2305.02220, 2023.

[11] Yu-Wen Chen and Julia Hirschberg. Exploring robustness in doctor-patient conversation summarization: An analysis of out-of-domain soap notes. arXiv preprint arXiv:2406.02826, 2024.

[12] Wei Chen, Zhiwei Li, Hongyi Fang, Qianyuan Yao, Cheng Zhong, Jianye Hao, Qi Zhang, Xuanjing Huang, Jiajie Peng, and Zhongyu Wei. A benchmark for automatic medical consultation system: frameworks, tasks and datasets. Bioinformatics, 39(1):btac817, 2023.

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

[14] Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. On llms-driven synthetic data generation, curation, and evaluation: A survey. arXiv preprint arXiv:2406.15126, 2024.

doi:10.6342/NTU202504251

[15] Huiyi Leong, Yifan Gao, Shuai Ji, Yang Zhang, and Uktu Pamuksuz. Efficient fine-tuning of large language models for automated medical documentation. In 2024 4th International Conference on Digital Society and Intelligent Systems (DSInS), pages 204–209. IEEE, 2024.

doi:10.6342/NTU202504251