MedSynth: Realistic, Synthetic Medical Dialogue-Note Pairs

Anonymous ACL submission

Abstract

Physicians spend significant time documenting clinical encounters, a burden that contributes to professional burnout. To address this, robust automation tools for medical documentation are crucial. We introduce MedSynth – a novel dataset of synthetic medical dialogues and notes designed to advance the Dialogue-to-Note (Dial-2-Note) and Note-to-Dialogue (Note-2-Dial) tasks. Informed by an extensive analysis of disease distributions, this dataset includes over 10,000 dialogue-note pairs covering over 2000 ICD-10 codes. We demonstrate that our dataset markedly enhances the performance of models in generating medical notes from dialogues, and dialogues from medical notes. The dataset provides a valuable resource in a field where open-access, privacy-compliant, and diverse training data are scarce. Code and data are available at https://anonymous.4open.science/r/MedSynth/.

1 Introduction

Electronic health records (EHRs) are widely used tools meant to streamline patient care (Melton et al., 2021). However, despite their benefits, EHRs also introduce a significant documentation burden on healthcare providers. Physicians report spending between 52 to 102 minutes daily on clinical note-taking from patient encounters, a task that is both time-consuming and detracting from direct patient care (Hripcsak et al., 2011). This extensive documentation requirement contributes significantly to physician burnout, highlighting an urgent need for solutions that can mitigate these demands while preserving the quality and integrity of patient records.

Automated tools for medical documentation have emerged as promising solutions to reduce this burden. However, developing and validating these tools is often hampered by the lack of large, open-access datasets that are comprehensive and privacy-compliant (Yim et al., 2023). Existing datasets are often limited in scope, cover only a few medical conditions, lack adherence to standard medical formats, or are generally unavailable due to privacy concerns. This restricts their utility and hampers progress (Wang et al., 2023; Yim et al., 2023).

Data augmentation offers an effective approach to bridge this gap. While synthetic data may not capture all nuances in real patient data, it can still provide valuable diversity and volume, enhancing ML training processes without compromising patient privacy. In this paper, we address the limitations of existing datasets by providing a novel synthetic dataset for data augmentation. Our primary contributions are:

- **A large synthetic medical dialogue-note dataset:** We provide a comprehensive, privacy-compliant dataset of medical dialogues paired with clinical notes encompassing a wide array of medical conditions over more than 2000 different ICD-10[1] codes and over 10,000 dialogue-note pairs. The notes in our dataset follow the SOAP structure (Subjective, Objective, Assessment, Plan), which is one of the most common formats of medical notes (Podder et al., 2024). The diversity of ICD-10 codes and the use of SOAP structure mimic the use case of primary care, which is an underrepresented specialty in AI research but generalizes to other specialties as well.

- **Novel data generation pipeline:** We introduce a novel pipeline for generating synthetic medical dialogue-note pairs, which is extensible to other domains.

- **SotA Dial-2-Note and Note-2Dial models:** We release models fine-tuned on our dataset and achieve SotA performance in generating medical notes from doctor-patient dialogues

---

[1]10th revision of the International Classification of Diseases (ICD) by the World Health Organization (WHO).

and vice versa. These models can be used as the base model for developing tools to automate documentation.

2 Related Work

Several datasets are commonly used in medical dialogue-to-note summarization. 3M Health (Zhang et al., 2021), Abridge (Krishna et al., 2020), Augmedix (Yim and Yetisgen-Yildiz, 2021), emr.ai (Finley et al., 2018), and Nuance (Enarvi et al., 2020) are important datasets for developing and testing models but their lack of public availability restricts their utility for reproducible research.

In contrast, open-source datasets such as MTS-dialogue (Abacha et al., 2023), PriMock57 (Korfiatis et al., 2022), Aci-Bench (Yim et al., 2023), and NoteChat (Wang et al., 2023) represent a meaningful advancement in improving the accessibility of medical dialogue data. Notably, some of these datasets, such as Aci-Bench, include synthetic dialogues created through human role-playing while others, like NoteChat (Wang et al., 2023), use large language models (LLMs) to generate dialogues.

However, these open datasets still have various limitations. They often concentrate on specific diseases — lacking a comprehensive range of medical conditions common in primary care – or they do not follow standard medical note structures. Table 1 shows a comparison of the available datasets.

3 Methodology

This section describes the methods taken to develop MedSynth, including an analysis of real-world disease distributions, important quality factors, and the novel data generation pipeline.

3.1 Data Resource

Four main data sources inform this research and are introduced in this section.

IQVIA PharMetrics Plus is a US medical insurance claims database, which has been the data source for several research findings such as cost savings associated with drug interactions (Yaghmaei et al., 2023). We use this database to estimate the frequency of ICD-10 codes. We generate data for the top 2000 most frequent medical conditions. We had access to a 25% random sample of the database from 2006 to 2021, which amounts to over 3 billion insurance claims. The database contains claims with either ICD-9 or ICD-10 code labels. ICD-10 codes were applied in practice in 2015

(Hirsch et al., 2016). Mapping ICD-9 codes to ICD-10 is not straightforward due to significant structural and conceptual differences between the two systems. ICD-10 is not simply a superset of ICD-9 — it introduces greater granularity (Cartwright, 2013) as well. Therefore, we filter the data to only include claims with ICD-10 codes. Moreover, according to Benes (2023), some codes are duplicated across the ICD-9 and ICD-10 systems. To avoid confusion, we remove the duplicates and end up with 800 million claim records and 54,985 unique ICD-10 codes.

The **Aci-Bench** dataset contains 207 dialogue-note pairs and was generated by medical experts and lay persons role-playing (Yim et al., 2023). As Table 1 presents, to the best of our knowledge, Aci-Bench is the largest publicly available dataset in terms of the length of the dialogues and notes that went through extensive human evaluation and is considered the latest benchmark for the medical Dial-2-Note and Note-2-Dial tasks and is publicly available under Creative Commons Attribution 4.0 International License.

NoteChat was introduced by Wang et al. (2023) and contains 207k dialogue-note pairs. The dialogues are synthetically generated by multiple cooperating role-playing LLMs. The notes come from PMC-Patient (Zhao et al., 2022) available under CC BY-NC-SA license, which reported a wide coverage of MeSH[2] disease terms. NoteChat is one of the largest publicly available datasets in this area, in terms of the number of data points. However, its notes are not provided in standard medical formats but are instead patient summaries (Zhao et al., 2022), which limits NoteChat's applicability for Dial-2-Note and Note-2-Dial tasks, as observed in our experiments.

PriMock57 contains 57 dialogue-note pairs generated through a simulation framework developed to mimic a telemedicine environment Korfiatis et al. (2022). Trained physicians and laypersons were included in the study to simulate doctor-patient encounters and, after each interaction, physicians wrote notes in the SOAP format.

3.2 Diseases Distribution Analysis

To ensure that our synthetic data reflects real-world disease prevalence, we analyze and identify the distributions of the most common diseases. To do so, we use the IQVIA PharMetrics Plus database.

---

[2]MeSH (Medical Subject Headings), used for indexing articles in PubMed.

Dataset	Description	Tasks	Src-len (tok/turns)	Target-len (tok/sent)	Size	Open
3M Health (Zhang et al., 2021)	Dialogue-note pairs: notes are created using dialogues	Dial-2-Note, Note-2-Dial	-/-	-/- (HPI only)	1342	No
Abridge (Krishna et al., 2020)	Dialogue-note pairs: notes are created using dialogues	Dial-2-Note, Note-2-Dial, ASR	1500/-	-/27	6862	No
Augmedix (Yim and Yetisgen-Yildiz, 2021)	Real clinical dialogue-note pairs	Clinical Dial-2-Note sentence alignment, Dial-2-Note	-/175	-/47	500	No
EMR.AI (Finley et al., 2018)	Real clinical dialogue-note pairs	Clinical form filling	616/1	550/-	9875	No
Nuance (Enarvi et al., 2020)	Real clinical dialogue-note pairs	Dial-2-Note, Note-2-Dial	972/-	452/-	802k	No
MTS-dialogue (Abacha et al., 2023)	Dialogue-note snippets: dialogues are created using clinical note sections	Dial-2-Note, Note-2-Dial	142/9	48/3	1701	Yes
PriMock57 (Korfiatis et al., 2022)	Dialogue-note pairs: role-played by human	Dial-2-Note, Note-2-Dial, ASR	1489/97	161/23	57	Yes
Aci-Bench (Yim et al., 2023)	Dialogue-note pairs: role-played by human	Dial-2-Note, Note-2-Dial	1,302/55	490/37	207	Yes
NoteChat (Wang et al., 2023)	Dialogue-note pairs: dialogue role-played by LLMs based on patient summaries	Dial-2-Note, Note-2-Dial	561/27	301/14	207k	Yes
MedSynth (this work)	Dialogue-note pairs: notes and dialogues role-played by LLMs	Dial-2-Note, Note-2-Dial	1070/47	621/23	10,035	Yes

Table 1: Comparison of Dial-2-Note generation datasets. 'ASR': Automatic Speech Recognition.

Table 2 shows that disease distribution is highly skewed. We choose the most frequent 2000 ICD-10 codes and generate 5 dialogue-note pairs for each. We use this uniform sampling, despite the data skew, to prevent MedSynth from being dominated by common diseases, as our goal is to generate a synthetic dataset that can be used to train models that are robust over a diverse set of diseases. To the best of our knowledge, our dataset is the largest dataset of medical dialogue-note pairs available publicly in terms of the number and diversity of diseases that adhere to standard medical structure. Table 14 in Appendix I shows the 20 most prevalent ICD-10 codes in our analysis.

# of Claims	# of ICD-10 Codes
Above 10M	4
1M - 10M	112
100k - 1M	1032
10k - 100k	3922
1k - 10k	8942
Below 1k	40,973

Table 2: Distribution of ICD-10 codes from IQVIA Pharmetrics Plus database.

3.3 Survey on Important Quality Factors

As mentioned by Mathioudakis et al. (2016), medical notes should adhere to standards to ensure continuity of care and facilitate communication between healthcare professionals. Good documentation should allow patients to access and understand the treatment procedure. Moreover, in case of an audit of healthcare services, medical notes can be a valuable resource. To ensure the synthetic medical notes capture useful information, particularly because real medical notes were inaccessible, we surveyed two medical professionals to determine the essential variables for note quality. One participant is a general physician and the other is a bio-statistician with over 13 years of experience with EHR systems and medical data. We listed 26 variables and asked the raters to score each variable between 0 and 10 in terms of how important each variable is in determining the quality of medical

notes. In the survey, 0 represents the lowest importance, and 10 represents the highest importance. We then take the 10 variables with the highest average importance scores. There was a tie in the 10th position, so we considered eleven variables initially. Appendix G discusses the results of the survey. Also, after obtaining the importance scores and discussing them with the medical experts, we added two more variables: Physical Exams and Investigations/Tests.

We use the results from this survey and the disease distribution analysis to guide our data generation pipelines.

3.4 Data Generation Pipeline

For our data generation, we first generate notes and then use these notes to generate the corresponding dialogues. This is due to the fact that it is relatively easier to define "quality" in medical notes compared to dialogues, and then we can generate dialogues based on medical notes. Multiple LLM agents collaborate to generate these notes. We use GPT-4o in both note generation and dialogue generation pipelines. We use two prompting strategies: Chain-of-Thought (CoT) (Wei et al., 2022) improves LLM performance in complex reasoning tasks through intermediate reasoning steps, and role-playing that acts as an effective trigger for the CoT process (Kong et al., 2023). All prompts are in Appendix C.

We also use In-Context Learning (ICL) as a training-free framework that does not require updating model parameters (Brown, 2020). ICL is directly applied to a pre-trained LLM and improves model performance by demonstrating good answers to queries without updating model parameters. In this paper, we use three forms of ICL:

a) Few-Shot ICL: At inference time, n demonstrations of the desired task are given to the model as conditioning in natural language.

b) One-Shot ICL: Similar to Few-Shot ICL, but with $n = 1$.

c) Zero-Shot ICL: Similar to Few-Shot ICL but a natural language description of the task is provided to the model instead of examples of the required task. We use this technique when there is no example available to provide to the model.

Previous research showed that model performance increases by increasing the number of examples in ICL (Agarwal et al., 2024; Brown, 2020). Therefore, we opt to use the maximum number of examples possible, while also considering com-
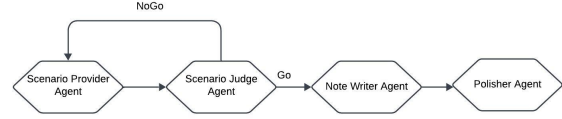

Figure 1: Note Generation Pipeline: The pipeline consists of four agents collaborating to generate the medical notes

putation costs. In this research, examples include dialogue-note pairs that translate to almost 1800 tokens per example. Therefore, we use three examples in the Few-Shot ICL setting

3.4.1 Note Generation Pipeline

Figure 1 shows that our note-generation pipeline that consists of four agents:

1) The **Scenario Provider Agent** receives the disease description (ICD-10 description), selects an appropriate role based on the disease (e.g., if the disease involves a broken bone, the Scenario Provider chooses Orthopedist), and generates a scenario based on the thirteen specified variables that fit the disease description and the chosen role. In each generation, we sample a note from the Aci-Bench training set and provide it to the model.

2) The **Scenario Judge Agent** is tasked with evaluating the quality and validity of scenarios generated by the Scenario Provider Agent. Its main responsibility is to determine whether the provided scenarios meet specific conditions to be approved for generating synthetic medical notes. The evaluation process involves three checks:

1. Compare variables: The agent checks whether at least four out of the thirteen variables in the current scenario differ from previously approved scenarios for the same ICD-10 code. The threshold of four was determined empirically as providing the balance between computational complexity and variety of the notes. This ensures diversity and uniqueness in the generated scenarios.

2. Medical Accuracy: The agent examines the scenario to confirm that it is medically accurate. This includes verifying that the symptoms, tests, diagnosis, and treatment are appropriate and logically consistent with the disease described by the ICD-10 code description.

3. Plausibility: The agent evaluates the overall plausibility of the scenario to ensure it is realistic and could feasibly occur in a medical

setting.

If a scenario is rejected, detailed feedback is provided by the Scenario Judge to the Scenario Provider. The feedback contains textual suggestions from the Scenario Judge about what parts of the scenario could be improved to pass the checks. Appendix H includes an example of such feedback. This loop continues until a pre-selected number of scenarios are approved.

3) The **Note Writer Agent** receives a detailed scenario and the physician's role (e.g., a general physician, or a specialist in various fields). Table 12 in Appendix E shows the frequency statistics of the roles selected by the agent. Based on this scenario and the role, the agent generates a medical note adhering to the SOAP format. Details of the SOAP format are provided in Appendix B. The Note Writer Agent uses One-Shot ICL; in each generation, we sample a note from the Aci-Bench training set and provide it to the model to help generate notes in a similar format.

4) The **d) Note Polisher Agent** is designed to refine and enhance synthetic medical notes. The primary task is to ensure that the information in the notes is correctly categorized and placed in the appropriate sections.

3.4.2 Dialogue Generation Pipeline

Our dialogue generation pipeline consists of two agents: the Dialogue Generator and the Dialogue Polisher.

a) The Dialogue Generator Agent generates a dialogue relevant to a given medical note. We use Few-Shot ICL with $n = 3$ and feed three examples from the Aci-Bench training set to boost the quality of the generation.

b) The Dialogue Polisher Agent enhances and expands doctor-patient conversations by including social chatter to make them more realistic. More importantly, it double-checks to ensure all information from the medical note is accurately included in the dialogue. This is done to reduce hallucinations in the generated dialogue.

4 Experiments

In this section, we analyze MedSynth's characteristics and assess the extrinsic quality of MedSynth by testing it on Dial-2-Note and Note-2-Dial tasks, demonstrating its effectiveness in training SotA models. All the experiments are run on a single A40 GPU and models are 4-bit quantized for computational efficiency.

4.1 Data Characteristics

Table 3 shows the statistical characteristics of MedSynth. It covers 2001 unique ICD-10 codes and over 10,000 dialogue-note pairs.

Metric	MedSynth
# of data points	10,035
# of unique ICD-10 codes	2001
Dialogue	
Avg length (tok)	932
Avg length (sentences)	55
Note	
Avg length (tok)	621
Avg length (sentences)	23

Table 3: Statistical characteristics of MedSynth. Note 'tok' does not include speaker IDs.

4.2 MedSynth for the Dial-2-Note and Note-2-Dial Generation

Human evaluation of generated data is costly and time-consuming (Liu et al., 2024). Therefore, we automatically assess the quality and usefulness of the data for Dial-2-Note and Note-2-Dial generation. Aci-Bench represents the latest benchmark for these tasks. The central hypothesis is that, if MedSynth's quality is high and its dialogues and notes closely resemble the distribution of Aci-Bench, then models fine-tuned on MedSynth should achieve strong performance on the Aci-Bench test set. We run a series of experiments to extrinsically evaluate the quality of MedSynth.

4.2.1 Evaluation Metrics

Recently research suggests that traditional metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) do not correlate well with human judgment in evaluating machine-generated text and, therefore, do not capture the depth and granularity of human evaluation (Zhang et al., 2019; Krishna et al., 2021). Using LLMs as a judges has become a prevalent alternative due to its potential parity with human evaluation (Zheng et al., 2023). Therefore, we use this setting as our primary evaluation strategy. We also report traditional metrics, i.e., ROUGE (Lin, 2004), BLEU (Papineni et al., 2002),

and METEOR (Banerjee and Lavie, 2005) in Appendix A.

Verga et al. (2024) reported that "juries" composed of diverse LLMs can outperform a single judge and decrease intra-model bias. Therefore, due to the sensitivity of medical applications such as our targeted tasks, we use majority voting with three LLM judges from different model families: Prometheus (Kim et al., 2023), GPT-4o, and Qwen2.5-32B-Instruct (Yang et al., 2024).

In our case, we provide the reference note, the notes generated by two models, and a rubric to each Judge LLM and ask it to decide which model performed best according to the rubric and reference note. Then, we use majority voting to decide which note is chosen. For the Dial-2-Note task, the rubric includes seven aspects: Hallucination, Critical Omissions, Professional Tone, Logical Structure of the Note and Sentences, Adherence to the SOAP Format, and Section Relevance. For the Note-2-Dial task, the rubric includes: Completeness, Accuracy, Naturalness and Flow, Use of Medical Terminology, and Evidence-Based Support. For both tasks, these criteria are selected based on previous research for the same tasks where researchers tried to gather human-expert feedback (Abacha et al., 2023; Wang et al., 2023). All the prompts and the detailed rubric can be found in Appendix C.

4.2.2 Results: Dial-2-Note Task

We sample data from NoteChat to match the sample size of MedSynth, combine these with the Aci-Bench training set, and fine-tune the Llama-3-8B-Instruct model. Finally, we test the tuned models on the Aci-Bench test set. To further evaluate the value of MedSynth, we run an additional experiment where the Aci-Bench training set is excluded from training to compare the standalone quality of MedSynth against NoteChat. We repeat the experiments to contrast MedSynth with PriMock57 instead of with NoteChat. Moreover, we contrast the model tuned on MedSynth and Aci-Bench's training set with the model solely trained on the Aci-Bench training set as a baseline. Table 4 shows the results of these experiments.

In all evaluations, when MedSynth is included in the training set, the corresponding model is preferred by the jury to any other model. We also observe that excluding the Aci-Bench training set from the training data widens the performance gap in favour of MedSynth. Results of the traditional

Comparison		#	Jury Pref. Rate
MS+AT vs. NC+AT		10,102	60.0%
MS Only vs. NC Only		10,035	95.0%
MS+AT vs. AT Only		10,102	52.5%
MS+AT vs. PM+AT		124	57.5%
MS Only vs. PM Only		57	90.0%

Table 4: LLM-based evaluation results on Aci-Bench for Dial-2-Note task. 'MS': MedSynth; 'AT': Aci_Train; 'PM': PriMock; 'NC': NoteChat; 'Jury Pref. Rate': Jury Preference Rate **in favour of MedSynth**. Note that the column **#** shows the number of samples in the training set.

metrics reported in Table **??** in Appendix A are also consistent with these observations.

An important observation is that when the models were fine-tuned exclusively on MedSynth compared with exclusively on NoteChat, a significant disparity in performance occurs in their adherence to the SOAP format. Specifically, the model fine-tuned only on NoteChat failed to maintain the structured SOAP format, which is a critical component of standard medical documentation. This confirms our expectation because NoteChat only includes patient summaries instead of structured medical notes. An example of such a failure is provided in Appendix F.

4.2.3 Results: Note-2-Dial Task

In the Note-2-Dial task, we use the same setting as the Dial-2-Note task and contrast MedSynth against NoteChat and PriMock57. We fine-tuned the Llama-3-8B-Instruct model under various conditions: one where the Aci-Bench training set was included alongside MedSynth and NoteChat, and another where it was excluded. The models were then evaluated on the Aci-Bench test set. We repeated the experiments with PriMock57 instead of NoteChat to further assess MedSynth's usefulness. Table 5 shows the results.

Similar to the Dial-2-Note task, the results show that including MedSynth in the training set leads to improved performance across all our experiments. Moreover, as observed in the Dial-2-Note task, excluding the Aci-Bench training set from training leads to bigger gaps in performance in favor of MedSynth. The results of the traditional metrics

presented in Table 9 in Appendix A are consistent with these observations.

Although MedSynth uses a simpler dialogue-generation pipeline compared to NoteChat, it outperforms NoteChat in generating dialogues from notes. This could be associated with the effectiveness of Few-Shot ICL in generating high-quality dialogues compared to the multi-agent setting without Few-Shot ICL used in NoteChat.

Comparison	#	Jury Pref. Rate
MS+AT vs. NC+AT	10,102	55.0%
MS Only vs. NC Only	10,035	87.5%
MS+AT vs. AT Only	10,102	80.0%
MS+AT vs. PM+AT	124	57.5%
MS Only vs. PM Only	57	97.5%

Table 5: LLM-based evaluation results for Note-2-Dial task. 'MS': MedSynth; 'AT': Aci_Train; 'NC': NoteChat; 'PM': PriMock; 'Jury. Pref. Rate': Jury Preference Rate **in favor of MedSynth.** Note that the column **#** shows the number of samples in the training set.

5   Ablation Studies

We conduct a series of ablation studies to assess aspects of our proposed pipeline. First, we ablated the Scenario Judge Agent by once removing it and once swapping the backbone LLM from GPT-4o to Llama-3.3-70B-Instruct. Second, we swapped all backbone LLMs in the whole data generation pipeline with either of two other models: once with Llama-3.3-70B-Instruct and once with Qwen2.5-32B-Instruct (Yang et al., 2024). We generated 246 samples in each of these configurations with the same ICD-10 codes. We repeat the experiments for both Dial-2-Note and Note-2-Dial tasks. All the experiments are run on a single A40 GPU and models are 4-bit quantized for computational efficiency.

5.1   Dial-2-Note Task

The results of the ablations for the Dial-2-Note task are presented in Table 6, and traditional metrics are reported in Table 10in Appendix A. By comparing the performance of the model tuned on the data generated by removing the Scenario Judge Agent from the pipeline (NuJudge+AT) with the one with

the agent present (AllGPT+AT), the first observation is that the Scenario Judge Agent is effective in the whole pipeline for the Dial-2-Note task. The win rate of AllGPT+AT over NuJudge+AT was 57.5%. However, the agent seems to be sensitive to the backbone LLM since swapping it to Llama-3.3-70B-Instruct (Lamma-JudgeGPT+AT) led to weaker performance. The win rate of Nu-Judge+AT over Lamma-JudgeGPT+AT was 52.5%. Traditional metrics point almost toward the same direction by showing lower scores for Lamma-JudgeGPT+AT. However, the scores were mostly the same for AllGPT+AT and NuJudge+AT.

The second observation is that not using GPT-4o as the backbone model decreased the performance, especially compared to NoteChat. It might be due to the fact that NoteChat was created using a multi-agent LLM collaboration, with the GPT family of models as the backbone. Further evaluation of using both pipelines with the same backbone family of models could shed more light.

Comparison	#	Jury Pref. Rate
NoJudge+AT vs. AllGPT+AT	313	57.5% **(AllGPT+AT)**
NoJudge+AT vs. LammaJudgeGPT+AT	313	52.5% **(NoJudge+AT)**
AllLamma+AT vs. PM+AT	124	52.5% **(AllLamma+AT)**
AllQwen+AT vs. PM+AT	124	52.5% **(PM+AT)**
AllLamma+AT vs. NC+AT	313	55.0% **(NC+AT)**
AllQwen+AT vs. NC+AT	313	60.0% **(NC+AT)**

Table 6: LLM-based evaluation results on Aci-Bench for ablation studies on Dial-2-Note task. 'MS': MedSynth; 'AT': Aci_Train; 'PM': PriMock; 'NC': NoteChat; 'Jury Pref. Rate': Jury Preference Rate (Winner). Note that the column **#** shows the number of samples in the training set.

5.2   Note-2-Dial Task

Table 7 shows that removing the Scenario Judge Agent and swapping the backbone model had a different effect on the Note-2-Dial task compared to the Dial-2-Note task. It suggests that the pipeline might be sensitive to the selection of the backbone model. By swapping all backbone models, we observe consistent performance changes with the Dial-2-Note task, suggesting that the pipeline seems to best perform with GPT-4o as the back-

bone. Further evaluation of using MedSynth's pipeline and NoteChat's pipeline with the same backbone family of models could be helpful.

Comparison	#	Jury Pref. Rate
NoJudge+AT vs. AllGPT+AT	313	55.0% **(NoJudge+AT)**
NoJudge+AT vs. LammaJudgeGPT+AT	313	55.0% **(LammaJudgeGPT+AT)**
AllLamma+AT vs. PM+AT	124	60.0% **(PM+AT)**
AllQwen+AT vs. PM+AT	124	55.0% **(AllQwen+AT)**
AllLamma+AT vs. NC+AT	313	67.5% **(NC+AT)**
AllQwen+AT vs. NC+AT	313	57.5% **(NC+AT)**

Table 7: LLM-based evaluation results on Aci-Bench for ablation studies on Note-2-Dial task. 'MS': MedSynth; 'AT': Aci_Train; 'PM': PriMock; 'NC': NoteChat; 'Jury Pref. Rate': Jury Preference Rate (Winner). Note that the column **#** shows the number of samples in the training set.

6 Discussion

We introduce MedSynth, a novel synthetic dataset designed to advance dialogue-to-note generation. MedSynth appears to be the first dataset of fully synthetic medical notes and dialogues, a crucial innovation in a field where data is scarce due to stringent privacy concerns. By incorporating over 2000 ICD-10 codes and more than 10,000 dialogue-note pairs, MedSynth overcomes the absence of diverse and privacy-compliant datasets for medical SOAP note generation. Although the focus of MedSynth is on primary care, it also contains diseases related to other specialties and can be used in any context that follows the SOAP format.

We show that models trained on MedSynth exhibit improvements in generating accurate and contextually appropriate medical notes from dialogues, which is essential for reducing the documentation burden on healthcare providers and addressing physician burnout. Existing datasets, such as Aci-Bench and NoteChat, though valuable, are limited either in the number of data points, the diversity of medical conditions covered, or in the (lack of) structure of medical notes. Aci-Bench, for instance, is limited in size with only 207 dialogue-note pairs and NoteChat uses patient summaries rather than structured medical notes.

We release our fine-tuned models that achieve SotA performance in the Dial-2-Note and Note-2-Dial tasks, which can be used as base models for future tool development.

7 Limitations and Ethical Considerations

First, while our pipeline is effective in generating synthetic data to improve Dial-2-Note and Note-2-Dial models, it does not guarantee the medical correctness of the generated text. Further research and expert evaluation are necessary to ensure medical accuracy. Therefore, MedSynth should not be considered a source of reliable medical information but rather a tool for model development.

Second, the synthetic nature of MedSynth, though beneficial for privacy, may not fully capture the complexities of real-world clinical dialogues between humans. Future work should evaluate the realism of these dialogues, to account for real-world sources of noise, and expand the dataset to include a broader range of diseases beyond the top 2000 ICD-10 codes, increasing its overall utility.

Third, although SOAP is among the most widely used structures for recording medical notes, other structures exist. Future work should consider covering more structures of medical notes.

Finally, while our results improve model performance in controlled settings, it is crucial to practice safe, careful 'MLOps' techniques when evaluating these models in clinical environments. Validating MedSynth's impact on documentation efficiency and accuracy in real-world clinical workflows, with 'real world' indicators of performance, will be an essential next step.

References

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. 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.

Rishabh Agarwal, Avi Singh, Lei M Zhang, Bernd Bohnet, Luis Rosias, Stephanie Chan, Biao Zhang, Ankesh Anand, Zaheer Abbas, Azade Nova, et al. 2024. Many-shot in-context learning. *arXiv preprint arXiv:2404.11018*.

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 ex-*

trinsic evaluation measures for machine translation and/or summarization, pages 65–72.

Gregory Benes. 2023. Risk of duplicate ICD codes for orthopedic and injury related research. *Perspectives in health information management*, 20(1).

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Donna J. Cartwright. 2013. ICD-9-CM to ICD-10-CM Codes: What? Why? How? *Advances in Wound Care*, 2(10):588–592. PMID: 24761333.

Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, Brian Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam McGrath, Yue Pan, Joel Pinto, et al. 2020. Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In *Proceedings of the first workshop on natural language processing for medical conversations*, pages 22–30.

Gregory Finley, Wael Salloum, Najmeh Sadoughi, Erik Edwards, Amanda Robinson, Nico Axtmann, Michael Brenndoerfer, Mark Miller, and David Suendermann-Oeft. 2018. From dictations to clinical reports using machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pages 121–128.

JA Hirsch, G Nicola, G McGinty, RW Liu, RM Barr, MD Chittle, and L Manchikanti. 2016. ICD-10: history and context. *American Journal of Neuroradiology*, 37(4):596–599.

George Hripcsak, David K Vawdrey, Matthew R Fred, and Susan B Bostwick. 2011. Use of electronic clinical documentation: time spent and team interactions. *Journal of the American Medical Informatics Association*, 18(2):112–117.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, et al. 2023. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations*.

Aobo Kong, Shiwan Zhao, Hao Chen, Qicheng Li, Yong Qin, Ruiqi Sun, Xin Zhou, Enzhi Wang, and Xiaohang Dong. 2023. Better zero-shot reasoning with role-play prompting. *arXiv preprint arXiv:2308.07702*.

Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163.

Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. Pri-Mock57: A dataset of primary care mock consultations. *arXiv preprint arXiv:2204.00333*.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. *arXiv preprint arXiv:2103.06332*.

Kundan Krishna, Sopan Khosla, Jeffrey P Bigham, and Zachary C Lipton. 2020. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. *arXiv preprint arXiv:2005.01795*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Lei Liu, Xiaoyan Yang, Fangzhou Li, Chenfei Chi, Yue Shen, Shiwei Lyu, Ming Zhang, Xiaowei Ma, Xiangguo Lv, Liya Ma, et al. 2024. Towards Automatic Evaluation for LLMs' Clinical Capabilities: Metric, Data, and Algorithm. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5466–5475.

Alexander Mathioudakis, Ilona Rousalova, Ane Aamli Gagnat, Neil Saad, and Georgia Hardavella. 2016. How to keep good clinical records. *Breathe*, 12(4):369–373.

Genevieve B Melton, Clement J McDonald, Paul C Tang, and George Hripcsak. 2021. Electronic health records. In *Biomedical Informatics: Computer Applications in Health Care and Biomedicine*, pages 467–509. Springer.

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.

V Podder, V Lew, and S Ghassemzadeh. 2022. SOAP Notes.[Updated 2022 Aug 29]. *StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing*.

Vivek Podder, Valerie Lew, and Sassan Ghassemzadeh. 2024. *SOAP Notes*, book chapter SOAP Notes. StatPearls Publishing, Treasure Island (FL). Available from StatPearls [Internet], PMID: 29489268, Copyright © 2024, StatPearls Publishing LLC.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models. *arXiv preprint arXiv:2404.18796*.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Ehsan Yaghmaei, Albert Pierce, Hongxia Lu, Yesha M Patel, Louis Ehwerhemuepha, Ahmad Rezaie, Seyed Ahmad Sajjadi, and Cyril Rakovski. 2023. A causal inference study: The impact of the combined administration of Donepezil and Memantine on decreasing hospital and emergency department visits of Alzheimer's disease patients. *PLoS One*, 18(9):e0291362.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

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(1):586.

Wen-wai Yim and Meliha Yetisgen-Yildiz. 2021. Towards automating medical scribing: Clinic visit dialogue2note sentence alignment and snippet summarization. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 10–20.

Longxiang Zhang, Renato Negrinho, Arindam Ghosh, Vasudevan Jagannathan, Hamid Reza Hassanzadeh, Thomas Schaaf, and Matthew R Gormley. 2021. Leveraging pretrained models for automatic summarization of doctor-patient conversations. *arXiv preprint arXiv:2109.12174*.

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

Zhengyun Zhao, Qiao Jin, Fangyuan Chen, Tuorui Peng, and Sheng Yu. 2022. Pmc-patients: A large-scale dataset of patient summaries and relations for benchmarking retrieval-based clinical decision support systems. *arXiv preprint arXiv:2202.13876*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P Xing, et al. 2023. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *arXiv preprint arXiv:2309.11998*.

A   Statistical Performance Metrics

This section reports the traditional statistical metrics used for model evaluation. All the metrics in this section are used from **evaluate** library from **hugggingface**.

Samples	Dataset	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-LSum	METEOR
67	AT	0.08 (0.05)	0.37 (0.14)	0.14 (0.07)	0.18 (0.05)	0.34 (0.13)	0.26 (0.08)
10102	MS+AT	0.14 (0.07)	0.54 (0.10)	0.24 (0.07)	0.26 (0.06)	0.51 (0.10)	0.33 (0.09)
10102	NC+AT	0.12 (0.06)	0.51 (0.11)	0.24 (0.06)	0.25 (0.07)	0.49 (0.10)	0.29 (0.07)
10035	MS	0.09 (0.04)	0.51 (0.07)	0.19 (0.03)	0.23 (0.04)	0.49 (0.06)	0.29 (0.06)
10035	NC	0.03 (0.02)	0.43 (0.08)	0.17 (0.04)	0.22 (0.05)	0.41 (0.08)	0.19 (0.05)
124	MS+AT	0.15 (0.07)	0.53 (0.09)	0.23 (0.05)	0.24 (0.05)	0.50 (0.09)	0.32 (0.07)
124	PM+AT	0.12 (0.07)	0.47 (0.13)	0.18 (0.07)	0.22 (0.06)	0.44 (0.13)	0.31 (0.08)
57	MS	0.09 (0.05)	0.51 (0.08)	0.19 (0.03)	0.23 (0.04)	0.49 (0.07)	0.30 (0.08)
57	PM	0.01 (0.01)	0.17 (0.13)	0.04 (0.04)	0.10 (0.07)	0.15 (0.12)	0.13 (0.10)

Table 9: Comparison of model performance by dataset and sample size for Note-2-Dial task. 'MS+AT': MedSynth+Aci_Train; 'NC+AT': NoteChat+Aci_Train; 'PM+AT': PriMock57+Aci_Train. Note that the column **Samples** shows the number of samples in the training set.

Samples	Dataset	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-LSum	METEOR
313	NoJudge+AT	0.26 (0.09)	0.59 (0.08)	0.32 (0.09)	0.38 (0.09)	0.55 (0.09)	0.45 (0.08)
313	AllGPT+AT	0.26 (0.09)	0.58 (0.11)	0.32 (0.09)	0.37 (0.10)	0.54 (0.11)	0.45 (0.09)
313	LammaJudgeGPT+AT	0.23 (0.09)	0.57 (0.09)	0.31 (0.09)	0.36 (0.09)	0.53 (0.09)	0.42 (0.09)
124	AllLamma+AT	0.18 (0.08)	0.50 (0.12)	0.25 (0.09)	0.30 (0.09)	0.46 (0.11)	0.39 (0.08)
124	AllQwen+AT	0.20 (0.07)	0.54 (0.09)	0.27 (0.07)	0.32 (0.08)	0.49 (0.09)	0.41 (0.06)
124	PM+AT	0.17 (0.07)	0.52 (0.10)	0.25 (0.07)	0.32 (0.08)	0.48 (0.09)	0.39 (0.08)
313	AllLamma+AT	0.23 (0.07)	0.57 (0.07)	0.31 (0.07)	0.36 (0.07)	0.53 (0.07)	0.43 (0.08)
313	AllQwen+AT	0.21 (0.10)	0.54 (0.15)	0.29 (0.11)	0.34 (0.11)	0.50 (0.14)	0.41 (0.11)
313	NC+AT	0.27 (0.10)	0.59 (0.10)	0.33 (0.09)	0.38 (0.10)	0.55 (0.10)	0.46 (0.10)

Table 10: Comparison of model performance by dataset and sample size for ablation studies on Dial-2-Note task. Note that the column **Samples** shows the number of samples in the training set.

Samples	Dataset	BLEU	ROUGE-1	ROUGE-2	ROUGE-L	ROUGE-LSum	METEOR
313	NoJudge+AT	0.13 (0.07)	0.52 (0.13)	0.23 (0.08)	0.25 (0.07)	0.48 (0.13)	0.30 (0.09)
313	AllGPT+AT	0.14 (0.06)	0.55 (0.10)	0.25 (0.07)	0.25 (0.06)	0.53 (0.10)	0.34 (0.07)
313	LammaJudgeGPT+AT	0.14 (0.07)	0.55 (0.10)	0.25 (0.06)	0.26 (0.06)	0.52 (0.09)	0.31 (0.07)
124	AllLamma+AT	0.10 (0.05)	0.45 (0.13)	0.18 (0.08)	0.22 (0.06)	0.42 (0.13)	0.28 (0.07)
124	AllQwen+AT	0.11 (0.06)	0.47 (0.13)	0.20 (0.07)	0.22 (0.06)	0.44 (0.13)	0.28 (0.08)
124	PM+AT	0.12 (0.07)	0.50 (0.11)	0.20 (0.07)	0.23 (0.06)	0.47 (0.11)	0.31 (0.09)
313	AllLamma+AT	0.11 (0.06)	0.45 (0.13)	0.18 (0.08)	0.22 (0.06)	0.42 (0.13)	0.31 (0.08)
313	AllQwen+AT	0.12 (0.06)	0.47 (0.13)	0.19 (0.08)	0.22 (0.06)	0.44 (0.13)	0.30 (0.07)
313	NC+AT	0.14 (0.07)	0.54 (0.09)	0.24 (0.06)	0.26 (0.06)	0.51 (0.09)	0.32 (0.08)

Table 11: Comparison of model performance by dataset and sample size for ablation studies on Note-2-Dial task. Note that the column **Samples** shows the number of samples in the training set.

B SOAP Format

The Subjective, Objective, Assessment, and Plan (SOAP) note is a standardized documentation framework widely used by healthcare professionals. This structured approach facilitates systematic and organized medical record-keeping, ensuring comprehensive and coherent documentation of patient encounters. We refer to Podder et al. (2022) for a more detailed explanation.

Subjective: The patient's description of their symptoms and complaints.

Objective: The physician's observations and collected data, including vital signs, physical examination findings, and test results.

Assessment: The physician's evaluation of the patient's condition, including the diagnosis or differential diagnosis.

Plan: The physician's recommendations for treatment, management, and follow-up.

C Prompts

This appendix contains the prompts we used in our experiments.

Assume you are a very experienced physician and you are conducting research. The research project is to generate synthetic medical notes from doctor-patient conversations. Your job is to provide a scenario for the note to be generated.
The notes must contain these variables:
** 1) Medical Outcome: Like diagnosis, prescribed treatment, follow-up recommendations, referral to specialists, referral to further tests or imaging, medication adjustment, lifestyle change (sleep, diet, exercise, tobacco use, alcohol use). If you think the note should contain prescribing medication, details must be included like dose, units, frequency, duration, quantity, quantity type (like tablets, etc.), and route (like oral or injected). If you think the note should contain referral, it must include details like the reason for the referral, the specialty, and the doctor's name. If you think the note should contain an order for blood work, it must include details like if it is for biochemistry, hematology, immunology, microbiology, viral hepatitis, vitamin D, prostate-specific antigen, or anything else that suits the scenario. You need to be very specific. If you think the note should contain an order for imaging, it must include details like the modality of the imaging and the area of the body. For example, if it is ultrasound, it can be an order for Abdominal, Thyroid, Musculoskeletal, Sonohysterogram, Sonohysterogram, Biophysical Profile(BPP), Scrotal, G.U. Tract - Kidneys-Bladder(Prostate), or anything else as it suits the scenario. Try to be very specific. REMEMEBER, you do not have to use all of them in the scenario. Use the ones that you see fit for the scenario.
** 2) Medical History: Can contain Previous Diagnoses, Family Medical History, Mdication History, Allergies, and Chronic Conditions.
** 3) Symptom Description: Can contain Severity, Duration, Associated Symptoms, Frequency, and Impact on Daily Activities.
** 4) Patient's self-reported habits and lifestyle: Can contain Sleep, Diet, Exercise, Tobacco Use, Alcohol Consumption, Drug Use, Recreational Activities.
** 5) Demographic Information: Can contain Age, Gender, Ethnicity, Socio-economic Status, Education Level, Health Literacy, and Job Status.
** 6) Patient's Behavior: The level of patient's cooperation with medical advice.
** 7) Geographical Location: Can contain Big City vs Small City, Rural vs Urban, Pollution and Environmental Health Risks, Neighborhood Type- eg if Impoverished or Affluent, Well-served by Transit, Food Desert, etc.
** 8) Clinical Setting: Can contain Hospitals, Clinics, Telemedicine, Mommunity Health Services, Urgent Care Centers, Research Facilities, School Health Services, Private Practice, and Specialty Clinics.
** 9) Type of Encounter: Can contain Initial Consultation, Follow-up, Emergency Visit, Routine Check-up, Chronic Disease Management (regular appointments to manage long-term health conditions like diabetes, heart disease, or chronic pain), Preventive Health Screening.
** 10) Treatment Disparities: There may be a tendency to offer less aggressive treatment or fewer options due to assumptions about compliance or ability to pay.
** 11) Native or Non-Native English Speaking Patient.
** 12) Physical exams: Any physical exams that are suitable for the scenario and the disease. If nothing is suitable, use NA.
** 13) Investigation/Test results: Like any tests that have been done for the patient while visiting. The results could be ready and reviewed in the scenario or could be awaiting. If awaiting, you need to be very specific about what type of tests have been done. For example, if it is X-ray, you need to incude the details mentioned abve about the imaging. You can also use NA if you think tests are not suitable for the scenario.
The user will give you the ICD-10 description of the disease. The diagnosis in the scenario must be the ICD-10 description. First, you select a role for yourself. You can be a Family Medicine Physician, a General physician, or a specialist with different specialties.
Select the role based on the ICD-10 description and output it with the keyword 'ROLE:'.
Second, you must come up with a scenario and list all the values for the variables you want to use in the scenario. Do not output any extra text, just your role at the top of the scenario and the list of the values. You should incorporate medication and blood work or imaging requests in the scenarios with the details mentioned above if it suits the scenario. These are artificial and people will not be using it without asking a real doctor.
The user will evaluate the scenario you provided. If they accept it, they do not give any feedback. If they do not accept your scenario, they will give you feedback about how to improve the scenario and you must incorporate the feedback and generate a new scenario.
Below is an example of a medical note. Remember, you need to provide the scenario, not the note.
EXAMPLE NOTE

Figure 2: Prompt for Scenario Provider LLM in Note Generation Pipeline

Assume you are a very experienced physician and you are conducting research. The research project is to generate synthetic medical notes from doctor-patient conversations. You have a coworker that works with you on the project. You have different roles. The coworker provides you with the scenario they are going to write notes with. Your job is to judge whether the scenario is approved or not based on the conditions I provided below. The scenarios have 13 variables.
** 1) Medical Outcome
** 2) Medical History
** 3) Symptom Description
** 4) Patient's self-reported habits and lifestyle
** 5) Demographic Information
** 6) Patient's Behavior
** 7) Geographical Location
** 8) Clinical Setting
** 9) Type of Encounter
** 10) Treatment Disparities
** 11) Native or Non-Native English Speaking Patient.
** 12) Physical exams: Any physical exams that are suitable for the scenario.
** 13) Investigation/Test results: Like any tests that have been done for the patient while visiting. The results could be ready and reviewed in the scenario or could be awaiting.
You have to check three conditions and then decide to approve or deny the scenario:
** a) A pair-wise comparison of the values of the variables. You need to check if at least 4 out of 13 of the values of the variables in the scenario are different from the previously approved scenarios.
** b) You need to check if the scenario is medically correct in terms of symptoms, tests, diagnosis, and treatment.
** c) You need to check if the scenario is plausible or not.
First, check condition (a). If it is not met, the scenario is rejected and you do not need to check conditions (b) and (c). If the scenario passes these three conditions, then say "Go". If not, you say "NoGo". In the case that there is no scenario previously approved, you should only check conditions (b) and (c). If your decision is "Go", you must only return "Go". Else, output your reasons and provide feedback to help your coworker about what they can do to generate a scenario to pass the above three conditions.
REMEMEBER: IF YOU APPROVE THE SCENARIO, YOU MUST ONLY OUTPUT LIKE BELOW. YOU CANNOT USE ANY BOLD OR ITALIC OR HEADING OR ANYTHING ELSE: DECISION: Go

Figure 3: Prompt for Scenario Judge LLM in Note Generation Pipeline

Assume you are a very experienced physician and you are conducting research.
The research project is to generate synthetic medical notes from doctor-patient conversations. The notes must be in this format:
** 1. Subjective: This section includes the patient's own description of their symptoms and complaints. Roll a dice, if the result is odd, break this part down into several sub-parts like Chief Complaint (CC), History of Present Illness (HPI), Review of Systems (ROS).
** 2. Objective: This section includes observations and data gathered by the physician, such as vital signs, physical examination findings, and test results.
** 3. Assessment: This section includes the physician's evaluation of the patient's condition, including a diagnosis or differential diagnosis.
** 4. Plan: This section includes the physician's recommendations for treatment, management, and follow-up.
You will be given a scenario containing your role. Your role can be a Family Medicine Physician, a General physician, or a specialist with different specialties. Your task is to generate the note based on the scenario. The note you generate must be in the format mentioned above.
Below is an example of a high quality medical note:
EXAMPLE NOTE

Figure 4: Prompt for Note Writer LLM in Note Generation Pipeline

Assume you are a very experienced physician and you are conducting research. The research project is to generate synthetic medical notes from doctor-patient conversations. The notes must be in this format:
** 1. Subjective: This section includes the patient's own description of their symptoms and complaints.
** 2. Objective: This section includes observations and data gathered by the physician, such as vital signs, physical examination findings, and test results.
** 3. Assessment: This section includes the physician's evaluation of the patient's condition, including a diagnosis or differential diagnosis.
** 4. Plan: This section includes the physician's recommendations for treatment, management, and follow-up.
You will be given a note. Your task is to polish the note and make sure the information is placed correctly in the relevant section. You cannot add or remove any information, except where you have been given permission. Make sure of these:
** a) If the doctor is ordering an imaging or bloodwork to be done, it must come under the "Plan" section. But if it is already done, it must come under the "Objective" section.
** b) If the doctor is prescribing a medication or renewing a medication, changing doses, etc., it must be under the "Plan" section.
** c) If the patient is being refered to another doctor, the referrals must come under the "Plan" section. The referral must contain:
1) The reason for referral
2) The specialty of the doctor
3) The doctor's name
If any of the three parts mentioned above about the refereal is missing, add it. If you need add a name for the doctor, choose an appropriate name, be creative and realistic in choosing the names.
** d) Patients' must have names. If there is no name, add it. Choose an appropriate name, be creative and realistic in choosing the names.
** e) The output must only contain the medical note. If there is anything extra at the beginning or at the end, you should remove it. For example, there could be thing like "note is: Medical Note" at the beggining or things like " Note: Please ensure this note is reviewed by the attending healthcare provider or physician for accuracy and completeness before being added to the patient's medical record." at the end of the medical note. Please remove them so that the output is only the medical note itself, not anything else.
Just output the revised note, not anything else.

Figure 5: Prompt for Note Polisher in Note Generation Pipeline

You are a helpful medical research assistance. You will be give a medical note and your task is to generate the conversation between the doctor and the patient that led to that note.

Your conversation must include all information. if it's difficult to include them all, you can use the original sentences in the notes. The common symptoms and common medical history should be told by patient.

Some specific symptoms and medical history should be added by the doctor after the patient has finished describing his symptoms and medical history.

For example:

Doctor: Can you give me your medical history record?
Patient: Here you are.
Doctor: Based on your medical history record...

Because after patient has finished describing common symptoms or medical history, he will give doctor his medical history records.

After patient give the doctor his medical history record, the doctor could know medical history record. Otherwise he didn't know any information of the medical history.

Some result should not come from history clinical note they should come from examination. All the examination result, history examination result, vital signs and medical number must be told by doctor.

You could expand the parts of doctor to include more key words. If it is difficult to include you could just use the sentence of clinical note. The revised conversation should be at least around 80 to 150 utterances(doctor or patient should not say too much information at once). The conversation must include all the information of the clinical note. You must include all the key words I gave you. If it is difficult to include all the key words you could use original the sentences of clinical note.

You cannot revise or eliminate any key words and you cannot use synonyms of the key words. You shoudn't use the abbreviation if you know the full name(you should use full name not abbreviation, such as D9 must be day 9, D7 must be day 7. If both the full name and the abbreviation appear, it's better to use the full name rather than the abbreviation.

Patients must not say any highly specialized terms, medical terminology or medical dosage. They can only describe limited common symptoms. The doctor should supplement the remaining information based on test results. Don't repeat the same information in long paragraphs. The utterance of the dialogue needs to be expanded as much as possible. The patient and the doctor should have many modal particles (e.g. hmm, yes, okay) to increase interaction. Pay attention to the examples below and try to incorporate non-linear discussions to make it more realistic.

You cannot use[Patient's Name] or any other plcae holder in the dialogue.

Here are a good real note and dialogue example:

Example 1:
Note:
EXAMPLE 1 NOTE
conversations:
EXAMPLE 1 DIALOGUE
««««««««««««««««««««««< Example 2:
Note:
EXAMPLE 2 NOTE
conversations:
EXAMPLE 2 DIALOGUE
«««««««««««««««««««««« Example 3:
Note:
EXAMPLE 3 NOTE
conversations:
EXAMPLE 3 DIALOGUE
««««««««««««««««««««««
You must follow the structure of the dialogues in the examples above.

The number of utterance should be at least 80 and sometimes patient didn't clearly hear and he could say parden to let the doctor say again. The dialogue must be in English. Your job is to only generate the dialogue. You cannot generate summary notes.

Figure 6: Prompt for Dialogue Generator in Dialogue Generation Pipeline

Expand the conversation. You must add chit chats to the conversation. The conversation for patient parts can be more colloquial. The patient and the doctor should have many modal particles (e.g. hmm, yes, okay) to increase interaction.

All the numbers and medical concepts that appear in the note should be mentioned by the doctor.

Professional medical terms and numbers should always occur in the doctor's utterances but not in the patient's answer. The doctor may describe and explain professional judgment to the patient and instruct the patient on follow-up requirements, but not ask questions that require professional medical knowledge to answer. All the information of medical history, symptoms and medication history should be told by patient.

The patient's answer should be succinct and accurate in a colloquial lay language style. The answer must align with the clinical notes and as colloquial as possible.

You can add some transitional phrases to make the conversation more logical. For example:

Example 1:
Patient: I understand, please go ahead.
(After examination)
Doctor: The result shows......

Your conversations can follow the logical sequence of a doctor's inquiry.

The conversations must be coherent and cohesive. For example, the output cannot be seperated by texts like "HISTORY OF PRESENT ILLNESS" or "SOCIAL HISTORY".

Extra information that does not fit into the conversation should not be added to the output. For example, below is an extra information that should be removed from the output:

««««<»»»»
- **INSTRUCTIONS**
Patient Agreements: The patient understands and agrees with the recommended medical treatment plan.
««««<»»»»

Patients should not say too much information at once.

ICD code of the disease must not be present in the dialogue. If it is present, remove it.

There should not be any extra information at the beggining or at the end of the conversation. For example, "dialogue is:" should not be present in the output. You must make sure you only return the dialogue itself, not anything extra. You cannot add phrases like "dialogue is: Certainly! Let's expand the conversation with more colloquial language for the patient and professional details for the doctor".

If there are only the doctor and the patient present in the dialogue, the utterances must follow these indicators:

[doctor]: ...
[patient]: ...

If there are more people present in the dialogue, make sure to include all of them with a seperate indicator. For example, if the mother of the patient is present in the dialogue, use this indicators:

[doctor]: ...
[mother]: ...
[patient]: ...

All the information in the dialogue must align with the medical note below:

««««<»»»»
MEDICAL NOTE
««««<»»»»

Figure 7: Prompt for Dialogue Polisher in Dialogue Generation Pipeline

1. Hallucination: - Does the summary note accurately and comprehensively reflect the doctor–patient dialogue and ground truth note?
2. Critical Omissions: - Does the summary note capture all essential medical facts from the doctor–patient dialogue and ground truth note?
3. Professional Tone: - Does the summary note maintain a consistently professional tone appropriate for expert use?
4. Logical Structure: - Does the summary note exhibit a clear and logical structure?
5. Adherence to the Format: - Does the summary note follow the same structure as the ground-truth note?
6. Section Relevance: - Does the summary note accurately assign clinical information to the correct sections (e.g., patient-reported details in Subjective, objective findings in Objective, clinician insights in Assessment, and treatment strategies in Plan)?

Figure 8: Evaluation rubric for Dialogue-to-Note evaluation

1. Completeness: - Does the conversation cover all significant components of the note? 2. Accuracy: - How accurately does the conversation reflect the details of the note as they were recorded? - Are there any discrepancies between the note and the generated dialogue? 3. Naturalness and Flow: - Is the conversation realistic and natural, following a logical and smooth progression? - Does it sound like a genuine interaction between a doctor and patient? 4. Use of Medical Terminology: - Is medical terminology used correctly and effectively within the conversational context? - Does the use of terminology enhance the accuracy and professionalism of the conversation? - Does the technical terminology used by the doctor and the patient represent their respective knowledge levels (e.g., layperson language for the patient)? 5. Evidence-Based Support: - Are the doctor's statements and responses consistent with the medical details and recommendations in the notes?

Figure 9: Evaluation rubric for Note-to-Dialogue evaluation

D Hyperparameters

This section contains details of hyperparameters in our experiments and the distribution of clinical note data lengths, by subtoken, for additional background.

```
Scenario Generator configuration in Note Generation Pipeline
scenario_generator_config = {
"model": "gpt-4o",
"temperature": 1,
"max_tokens": 4000,
"top_p": 1,
"frequency_penalty": 0,
"presence_penalty": 0,
}
```

Figure 10: Configuration for Scenario Generator in the Note Generation Pipeline

```
Scenario Judge configuration in the Note Generation Pipeline
scenario_judge_config = {
"model": "gpt-4o",
"temperature": 0,
"max_tokens": 4000,
"top_p": 1,
"frequency_penalty": 0,
"presence_penalty": 0,
}
```

Figure 11: Configuration for Scenario Judge in the Note Generation Pipeline

```
Note Generator configuration in the Note Generation Pipeline
note_generator_config = {
"model": "gpt-4o",
"temperature": 0.9,
"max_tokens": 4000,
"top_p": 1,
"frequency_penalty": 0,
"presence_penalty": 0,
}
```

Figure 12: Configuration for Note Generator in the Note Generation Pipeline

```
Note Polisher Configuration in the Note Generation Pipeline
note_polisher_config = {
"model": "gpt-4o",
"temperature": 0,
"max_tokens": 4000,
"top_p": 1,
"frequency_penalty": 0,
"presence_penalty": 0,
}
```

Figure 13: Configuration for Note Polisher in the Note Generation Pipeline

```
Dialogue Generator Configuration in the Dialogue Generation Pipeline
dialogue_generator_config = {
"model": "gpt-4o",
"temperature": 0.7,
"max_tokens": 4095,
"top_p": 1,
"frequency_penalty": 0,
"presence_penalty": 0,
}
```

Figure 14: Configuration for Dialogue Generator in the Dialogue Generation Pipeline

```
Dialogue Polisher configuration in the Dialogue Generation Pipeline
dialogue_polisher_config = {
"model": "gpt-4o",
"temperature": 0.5,
"max_tokens": 4095,
"top_p": 1,
"frequency_penalty": 0,
"presence_penalty": 0,
}
```

Figure 15: Configuration for Dialogue Polisher in the Dialogue Generation Pipeline

Defining the configuration for the base model, LoRA and training

```
tuning_config = {
"model_config": {
"max_seq_length": 8192,
"dtype": torch.bfloat16,
"load_in_4bit": True,
},
"lora_config": {
"r": 16,
"target_modules": ["q_proj", "k_proj",
"v_proj", "o_proj", "gate_proj",
"up_proj", "down_proj"],
"lora_alpha": 16,
"lora_dropout": 0,
"bias": "none",
"use_gradient_checkpointing": True,
"use_rslora": False,
"use_dora": False,
"loftq_config": None,
},
"training_config": {
"per_device_train_batch_size": 2,
"gradient_accumulation_steps": 4,
"warmup_steps": 5,
"max_steps": 0,
"num_train_epochs": 4,
"learning_rate": 2e-4,
"fp16":                               not
torch.cuda.is_bf16_supported(),
"bf16": torch.cuda.is_bf16_supported(),
"logging_steps": 1,
"optim": "adamw_8bit",
"weight_decay": 0.01,
"lr_scheduler_type": "linear",
"seed": 42,
"output_dir": "outputs"
}
}
```

Figure 16: Configuration for the base model, LoRA, and training setup in tuning models for evaluating the usefulness of MedSynth for the Dial-2-Note and Note-2-Dial tasks; Unsloth library was used to achieve higher speed

Updated Tuning Configuration

```
tuning_config = {
"model_config": {
"max_seq_length": 8192,
"dtype": torch.bfloat16,
"load_in_4bit": True,
},
"lora_config": {
"r": 16,
"target_modules": ["q_proj", "k_proj",
"v_proj", "o_proj", "gate_proj",
"up_proj", "down_proj"],
"lora_alpha": 16,
"lora_dropout": 0,
"bias": "none",
"use_gradient_checkpointing": True,
"use_rslora": False,
"use_dora": False,
"loftq_config": None,
},
"training_config": {
"per_device_train_batch_size": 2,
"gradient_accumulation_steps": 4,
"warmup_steps": 5,
"max_steps": 0,
"num_train_epochs": 1,
"learning_rate": 2e-4,
"fp16":                               not
torch.cuda.is_bf16_supported(),
"bf16": torch.cuda.is_bf16_supported(),
"logging_steps": 1,
"optim": "adamw_8bit",
"weight_decay": 0.01,
"lr_scheduler_type": "linear",
"seed": 42,
"output_dir": "outputs"
}
}
```

Figure 17: Configuration for the base model and LoRA training for tuning the models used in benchmarking; Unsloth library was used to achieve higher speed

```
Model evaluator generation configuration
config = {
"max_new_tokens": 3000,
"do_sample": True,
"temperature": 0.6,
"top_p": 0.9,
"use_cache": True,
}
```

Figure 18: Configuration for inference of tuned models in tuning models for evaluating the usefulness of MedSynth for the Dial-2-Note and Note-2-Dial tasks; Unsloth library was used to achieve higher speed

```
GPT-4o model configuration used in benchmarking
config = {
"model": "gpt-4o",
"temperature": 0.5,
"max_tokens": 4000,
"top_p": 1,
}
```

Figure 19: GPT-4o model configuration used in benchmarking

```
Inference configuration of tuned models used in benchmarking
config = {
"max_new_tokens": 4000,
"do_sample": True,
"temperature": 0.5,
"top_p": 1,
"use_cache": True,
}
```

Figure 20: Inference configuration of tuned models used in benchmarking; Unsloth library was used to achieve higher speed

792
793
794
795
796
797
798
799

800

801
802
803
804
805
806
807
808
809

E   Details about Roles in Scenario Provider Agent

The analysis of the generated scenarios shows that there are 258 unique roles in the whole dataset. Family Medicine Physician is the most common role, followed by Orthopedic Specialist, Neurologist, and General Physician. Table 12 shows the distribution of roles, and Table 13 shows the top 20 roles.

# of Occurrences	Count
More than 1000	1
Between 500 and 1000	4
Between 100 and 500	18
Between 50 and 100	6
Between 10 and 50	26
Below 10	203

Table 12: Frequency distribution of roles in Scenario Provider Agent.

Role	Frequency
Family Medicine Physician	1960
Orthopedic Specialist	613
Neurologist	584
General Physician	556
Gastroenterologist	526
Cardiologist	490
Dermatologist	371
Pulmonologist	326
Oncologist	326
Endocrinologist	305
Ophthalmologist	296
Obstetrician	289
Urologist	283
Orthopedic Surgeon	233
General Surgeon	221
Gynecologist	196
Psychiatrist	193
Nephrologist	166
Hematologist	159
Rheumatologist	153

Table 13: Top 20 most frequent roles in Scenario Provider Agent.

F   Comparison of Notes Generated by Models Trained on MedSynth and NoteChat

When tuned solely on NoteChat, the model failed to adhere to the SOAP structure in generating medical notes from dialogues. In the same setting, the model tuned on MedSynth successfully generated notes in SOAP format. This appendix shows an example of such a situation.

CHIEF COMPLAINT
Rheumatoid arthritis management.
HISTORY OF PRESENT ILLNESS
Mr. Nelson is a 72-year-old male who presents today for ongoing management of rheumatoid arthritis. He was last seen by me on 03/04/2021, where he received a steroid injection to his right knee. He continues to experience right knee pain and swelling. The pain level is 6/10. He notes when he stands after he has been sitting for a while, he has to use a cane to "get his bearings." The injection he received to the right knee at the last visit was only beneficial for a few weeks. The patient notes whole right leg swelling since the last visit. He usually wears support hose, which helps to some degree. He has not attended physical therapy.
PHYSICAL EXAM
Musculoskeletal Lower right extremity: Swollen right knee, slightly reduced range of motion. Skin warm to touch. Lower right leg pits with pressure.
PROCEDURE
Fluid aspiration and Steroid injection. Risks, benefits, and alternatives were explained. The risks of steroid injection were explained, including but not limited to, hypopigmentation, soft tissue atrophy, tendon ruptures, and transient hyperglycemia among others. Verbal consent was obtained. The right knee was accessed through the lateral approach with 3 mL of initially clear fluid then blood-tinged fluid. Then the right knee was prepped in a standard fashion. The right knee was injected with 80 mg triamcinolone and 1 mL lidocaine. The procedure was tolerated well with no complications. Post procedure instructions were given.
ASSESSMENT
• Rheumatoid arthritis management. • High risk medication use. • Right knee pain. • Lower extremity pitting edema.
PLAN
Rheumatoid arthritis management. The patient will continue Xeljanz XR 11 mg daily. He will also continue Celebrex 200 mg daily.
High-risk medication use. The patient is currently on the high-risk medication Xeljanz. We will check his CMP levels and monitor periodically every couple of months.
Right knee pain. He has right knee pain and swelling, with pain level 6/10. On exam today, the right knee continues to be hot, which is ongoing inflammation. We performed a fluid aspiration of the right knee today, see procedure note above. A steroid injection to the right knee was given today, see procedure note above. We will arrange for x-ray of the right knee and refer his to physical therapy.
Lower extremity pitting edema. The patient notes whole right leg swelling since the last visit. He usually wears support hose, which helps to some degree. He has not seen his PCP. We discussed compression stockings, elevate feet, and a low salt diet. He will check with his PCP for further instructions.

Figure 21: Ground Truth Note

Tyler Nelson, a 70-year-old male with a date of birth of 3/6/1949, social security number 385-729- and a home address of [redacted] presented to the office for ongoing management of rheumatoid arthritis. He received a steroid injection in the right knee last time he was in the office. He stated that the injection helped for a couple of weeks, it's just it. He stated that his right knee is swollen, and it's painful. He stated that his pain is a 3 or so on the pain scale, typically on a normal day it's around a 6. It's been swollen like that for how long now? He stated that it's been about two months now. Before it was just the knee was not swollen like this, but he stated that he tried to wear compression socks a lot of the time of the leg, but it has not been helping. He stated that he slowed down a little bit, and he found a few things too fast. When it really hurts is when he's been sitting for a while, and then he's got to try and get up. He's so stiff getting up he's got to use a cane to get his bearings. He means that after a little bit, it loosens up, but that first getting up is real difficult. He is on xeljanz XR 11 milligrams a day, he's taking it everyday, and he's also taking celebrex which he started maybe 2 months ago. The patient sat down and removed his socks and shoes. The right leg was swollen and hot. The patient stated that it hurts when he tries to straighten it all the way. Range of motion slightly diminished. The patient stated that the whole thing is swollen, not just the knee. There was some pitting around the knee and ankle. Normal capillary refill in the toes. Does any of this area hurt? The patient stated that it's kinda sore, like an old bruise? The patient agreed to get an x-ray of the knee. The patient was called for physical therapy a couple of years ago for the other knee. So, it was set up for him. The patient stated that it's been helpful to some degree, but it's still swollen, especially around the knee. The patient stated that he's got to try and get up, and it's real difficult. The patient was told that we'll set that up for him now. We'll also do an x-ray on that knee. We'll also set him for a physical therapy, and if nothing is helping then we really might need to get a surgery consult to see what we can do with that knee. The patient agreed to go ahead and we gave him the steroid shot. We tried to get a little bit of fluid out before we gave him the injection, there was no fluid but we saw if we could get anything out regardless. The procedure was explained, the risks, benefits, and alternatives were explained. The risks of steroid injection were explained, including, but not limited to: hypopigmentation, soft tissue atrophy, tendon ruptures, and transient hypoglycemia among others. Verbal consent was attained. The right knee was, uh, um, accessed through the lateral approach with 3 milliliters of initially clear fluid, then bloo- blood tinged fluid. Then the right knee was prepped in a standard-fashion. The right knee was injected with 80 milligrams of triamcinolone, and one milliliter of lidocaine. The procedure was tolerated well, with no complications. Post-procedure instructions were given. Patient will follow up with his pcp. Thank you.

Figure 22: Note Generated by the Model Trained Solely on NoteChat

1. Subjective:
Chief Complaint (CC): Swelling and pain in the right knee.
History of Present Illness (HPI): Tyler Nelson is a 73-year-old male with a history of rheumatoid arthritis who presents for a follow-up visit regarding the management of his condition. He reports that he experienced significant swelling in his right knee following a recent steroid injection administered during his last visit. The swelling has persisted despite his use of compression socks, which he has been wearing frequently to alleviate symptoms. The pain is moderate, with a pain score of 6/10 on a normal day and currently 3/10, and has been present for the past 2 months. He notes that the swelling and pain have limited his mobility, particularly when getting up from sitting, where he requires the use of a cane.
He has been compliant with his medication regimen, including Xeljanz 11mg daily and Celebrex 200mg daily, which he started 2 months ago. He is motivated to follow medical advice and is highly cooperative.
Review of Systems (ROS): - **Musculoskeletal:** Swelling and pain in the right knee, moderate severity, limiting mobility. - **General:** No recent weight loss, fever, or chills. - **Cardiovascular:** No chest pain or palpitations. - **Gastrointestinal:** No nausea, vomiting, or abdominal pain. - **Neurological:** No headaches or dizziness. - **Dermatological:** No rashes or skin changes.
2. Objective:
Physical Examination: - **Vital Signs:** Stable - **Musculoskeletal:** - Right knee: Swelling, erythema, and pitting noted. Range of motion slightly diminished. Increased heat around the knee. Capillary refill within normal limits in the toes. - **Cardiovascular:** Regular rate and rhythm, no murmurs. - **Respiratory:** Clear to auscultation bilaterally. - **Neurological:** Alert and oriented, no focal deficits.
Results: - **Blood Work:** CMP within normal limits. Xeljanz trough levels checked; levels are within the therapeutic range.
3. Assessment:
1. **Rheumatoid Arthritis with right knee effusion, not elsewhere classified.** - **Medical Reasoning:** The patient presents with persistent swelling and moderate pain in the right knee, consistent with effusion. The swelling has limited his mobility and daily activities. - **Differential Diagnosis:** Septic arthritis, crystal-induced arthritis, gout, lymphedema.
4. Plan:
1. **Treatment:** - Administered steroid injection (Triamcinolone 80mg, 1mL) to the right knee today. - Fluid aspiration performed during the procedure; 3mL of initially clear fluid, then blood-tinged fluid was aspirated.
2. **Imaging:** - Order an X-ray of the right knee to evaluate for possible underlying causes of effusion.
3. **Physical Therapy:** - Referral for physical therapy to aid in improving range of motion and strength.
4. **Follow-up Recommendations:** - Follow-up visit in 2 weeks for re-evaluation of the knee. - Referral to Dr. Emily Carter, a rheumatologist, for further management of rheumatoid arthritis.
5. **Patient Education and Counseling:** - Discussed the importance of reducing salt intake to help manage edema. - Encouraged the patient to elevate the legs when possible to reduce swelling. - Advised the patient to continue using compression socks as a preventive measure. - Discussed the potential need for surgical intervention if conservative measures fail to improve symptoms.
6. **Medication Adjustment:** - Continue Xeljanz 11mg daily. - Continue Celebrex 200mg daily. - Regular monitoring of Xeljanz trough levels every 2 months due to the high-risk medication profile.
Patient Agreements: The patient understands and agrees with the recommended medical treatment plan.
INSTRUCTIONS: The patient will follow up in 2 weeks for re-evaluation and will have an X-ray of the right knee performed. He will also make an appointment with Dr. Emily Carter, a rheumatologist, and ensure regular blood work for Xeljanz trough levels.

Figure 23: Note Generated by the Model Trained Solely on MedSynth

G Survey on Important Quality Factors

Figure 24 shows the results of our survey on the important variables in determining the quality of medical notes.
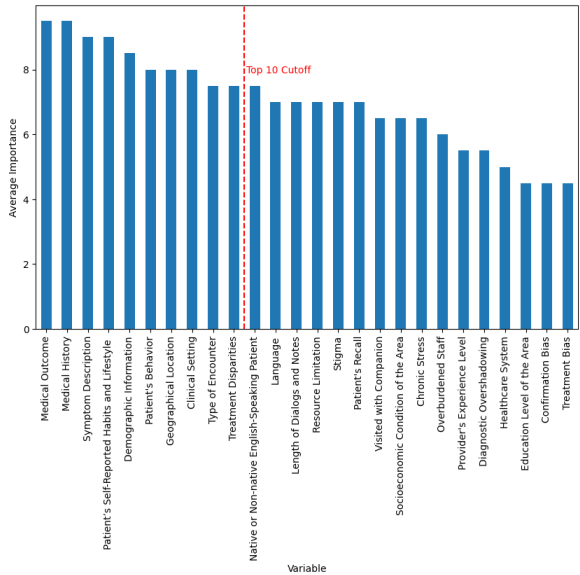

Figure 24: Results of the medical notes quality factors

We calculated Intraclass Correlation Coefficient (ICC) (Koo and Li, 2016) to measure the agreement between the annotators. There are multiple forms of ICC. Given the characteristics of our survey and the guide from Koo and Li (2016), we used $ICC(2, k)$. The observed $ICC(2, k)$ score was 0.56, which is considered a moderate score.

H Sample of Notes and Feedback in Data Generation Pipeline

I Data Statistics Tables

DECISION: NoGo
Reasons:
1. Condition (a) is not met: The scenario does not differ sufficiently from the previously approved scenario. The variables such as Medical Outcome, Symptom Description, Patient's Behavior, Geographical Location, Clinical Setting, and Type of Encounter are very similar to the previously approved scenario. There are not enough differences in at least 4 out of the 13 variables.
Feedback:
- Consider varying the demographic information, such as age, gender, or ethnicity, to create a more distinct scenario.
- Introduce a different medical history or chronic condition to differentiate the patient's background.
- Alter the geographical location or clinical setting to provide a new context for the scenario.
- Modify the patient's self-reported habits and lifestyle to reflect a different lifestyle or behavior pattern.
- Explore different types of encounters or treatment disparities to add variety to the scenario.

Figure 25: An example of feedback provided from Scenario Judge Agent to Scenario Provider Agent

1. Subjective:
Chief Complaint (CC):
Moderate wheezing and shortness of breath.
History of Present Illness (HPI): The patient, Maria Gonzalez, is a 35-year-old Hispanic female with a known history of asthma who presents with worsening symptoms over the past three days. She reports moderate wheezing, shortness of breath, chest tightness, and occasional coughing. Symptoms are exacerbated at night and with physical exertion, significantly impacting her daily activities and sleep.
Review of Systems (ROS):
- **Constitutional:** Reports interrupted sleep averaging 5-6 hours per night due to asthma symptoms.
- **Respiratory:** Positive for wheezing, shortness of breath, chest tightness, and occasional coughing.
- **Cardiovascular:** Denies chest pain or palpitations.
- **Gastrointestinal:** Denies nausea, vomiting, or changes in bowel habits.
- **Musculoskeletal:** Denies joint pain or swelling.
- **Neurological:** Denies headaches, dizziness, or syncope. **2. Objective:** **Vital Signs:**
- Temperature: 98.6°F - Pulse: 88 bpm - Respiratory Rate: 22 breaths per minute - Blood Pressure: 120/80 mmHg
- Pulse Oxygenation: 95% on room air **Physical Examination:**
- **General:** Patient appears in moderate respiratory distress, no cyanosis.
- **Respiratory:** Bilateral wheezing on auscultation, decreased breath sounds in the lower lobes.
- **Cardiovascular:** Regular rate and rhythm; no murmurs, gallops, or rubs.
- **Abdomen:** Soft, non-tender, no hepatosplenomegaly.
- **Extremities:** No clubbing, cyanosis, or edema.
3. Assessment:
Primary Diagnosis: Mild persistent asthma with (acute) exacerbation.
Differential Diagnoses:
- Chronic obstructive pulmonary disease (COPD)
- Respiratory tract infection
- Cardiac-related dyspnea
4. Plan:
Medical Treatment:
- **Albuterol Inhaler 90 mcg:** 2 puffs every 4-6 hours as needed for wheezing or shortness of breath.
- **Fluticasone Propionate Inhaler 110 mcg:** 2 puffs twice daily. **Follow-up and Referrals:**
- Schedule follow-up appointment in 2 weeks.
- Referral to Pulmonologist, Dr. Amanda Smith, for further asthma management.
Investigations:
- Spirometry to evaluate lung function.
- Complete Blood Count (CBC) and Allergy Panel to evaluate potential allergens and eosinophilia.
Lifestyle and Patient Education:
- Encourage avoidance of known allergens.
- Emphasize the importance of regular medication adherence.
- Discuss and implement an asthma action plan.
- Instruct on proper inhaler technique.
Patient Agreements:
The patient understands and agrees with the recommended medical treatment plan and follow-up. She is highly cooperative and adherent to medical advice.

Figure 26: An example of SOAP note generated by the Note Generation pipeline

ICD-10 Description	Frequency
ESSENTIAL (PRIMARY) HYPERTENSION	21,788,529
ENCOUNTER FOR GENERAL ADULT MEDICAL EXAMINATION WITH-OUT ABNORMAL FINDINGS	19,497,471
ENCOUNTER FOR ROUTINE CHILD HEALTH EXAMINATION WITH-OUT ABNORMAL FINDINGS	13,272,519
ENCOUNTER FOR IMMUNIZATION	10,470,154
TYPE 2 DIABETES MELLITUS WITHOUT COMPLICATIONS	9,890,112
END STAGE RENAL DISEASE	9,040,289
HYPERLIPIDEMIA, UNSPECIFIED	8,061,616
LOW BACK PAIN	7,336,799
OBSTRUCTIVE SLEEP APNEA (ADULT) (PEDIATRIC)	7,073,161
ILLNESS, UNSPECIFIED	6,253,356
HYPOTHYROIDISM, UNSPECIFIED	5,618,816
ENCOUNTER FOR SCREENING MAMMOGRAM FOR MALIGNANT NEOPLASM OF BREAST	5,393,554
ENCOUNTER FOR GYNECOLOGICAL EXAMINATION (GENERAL) (ROUTINE) WITHOUT ABNORMAL FINDINGS	5,299,912
URINARY TRACT INFECTION, SITE NOT SPECIFIED	5,075,284
OTHER LONG TERM (CURRENT) DRUG THERAPY	4,724,627
CERVICALGIA	4,621,704
ATHEROSCLEROTIC HEART DISEASE OF NATIVE CORONARY ARTERY WITHOUT ANGINA PECTORIS	4,504,152
CHEST PAIN, UNSPECIFIED	4,396,428
GASTRO-ESOPHAGEAL REFLUX DISEASE WITHOUT ESOPHAGITIS	4,224,803
ACUTE KIDNEY FAILURE, UNSPECIFIED	3,774,237

Table 14: Top 20 most frequent ICD-10 codes in IQVIA.