scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Aci-bench: a Novel Ambient Clinical Intelligence Dataset for Benchmarking Automatic Visit Note Generation

Wen-wai Yim[1]✉, Yujuan Fu[2], Asma Ben Abacha[1], Neal Snider[3], Thomas Lin[1] & Meliha Yetisgen[2]

Recent immense breakthroughs in generative models such as in GPT4 have precipitated re-imagined ubiquitous usage of these models in all applications. One area that can benefit by improvements in artificial intelligence (AI) is healthcare. The note generation task from doctor-patient encounters, and its associated electronic medical record documentation, is one of the most arduous time-consuming tasks for physicians. It is also a natural prime potential beneficiary to advances in generative models. However with such advances, benchmarking is more critical than ever. Whether studying model weaknesses or developing new evaluation metrics, shared open datasets are an imperative part of understanding the current state-of-the-art. Unfortunately as clinic encounter conversations are not routinely recorded and are difficult to ethically share due to patient confidentiality, there are no sufficiently large clinic dialogue-note datasets to benchmark this task. Here we present the Ambient Clinical Intelligence Benchmark (ACI-BENCH) corpus, the largest dataset to date tackling the problem of AI-assisted note generation from visit dialogue. We also present the benchmark performances of several common state-of-the-art approaches.

Background & Summary

Healthcare needs are an inescapable facet of daily life. Current patient care at the medical facilities requires involvement not only from a primary care provider, but also from pharmacy, billing, imagining, labs, and specialist care. For every encounter, a clinical note is created as documentation of clinician-patient discussions, patient medical conditions. They serve as a vital record for clinical care and communication with patients and other members of the care team, as well as outline future plans, tests, and treatments. Similar to typical meeting summaries, these documents should highlight important points while compressing itemized instances into condensed themes; unlike typical meeting summaries, clinical notes are purposely and technically structured into semi-structured documents, contain telegraphic and bullet-point phrases, use medical jargon that do not appear in the original conversation, and will reference outside information often from the electronic medical record, including prose-written content or injections of structured data.

While the widespread adoption of electronic health records (EHR's), spurred by the HITECH Act of 2009, has led to greater health information availability and interoperability, it has also spawned a massive documentation burden shifted to clinicians. Physicians have expressed concerns that writing notes in electronic health records (EHRs) takes more time than using traditional paper or dictation methods. As a result, notes may not be completed and accessible to other team members until long after rounds[1]. Furthermore, as another unintended consequence of EHR use complications, electronic notes have been criticized for their poor readability, completeness, and excessive use of copy and paste[2]. To save time and adequately capture details, clinicians may choose to write their notes during their time with a patient. This may detract from the clinicians' attention toward the patient (e.g. in reading non-verbal cues), and may leave patients feeling a want of empathy[3]. Alternatively, some clinicians or provider systems may hire medical assistants or scribes to partake in some

[1]Microsoft, Health AI, Redmond, 98052, USA. [2]University of Washington, Biomedical and Health Informatics, Seattle, 98109, USA. [3]Nuance Communications, Healthcare R&D, Burlington, 01803, USA. ✉e-mail: yimwenwai@microsoft.com

dataset	description	src-len (tok/turns)	target-len (tok/sent)	size	open
MTS-dialogue[13]	dialogue-note snippets where conversations are created using clinical note sections	142/9	48/3	1701	Y
primock57[14]	role-played dialogue-note pairs	1489/97	161/23	57	Y
ACI-BENCH [this work]	role-played dialogue-note pairs	1302/55	490/49	207	Y
3 M Health[9]	dialogue-note pairs where notes are created using conversations	–/–	–/– (hpi only)	1342	N
Abridge[8]	dialogue-note pairs where notes are created using conversations	1500/–	–/27	6862	N
Augmedix[11]	real clinical dialogue-note pairs	–/175	–/47	500	N
emr.ai[6]	real clinical dictation-note pairs	616/1	550/-	9875	N
Nuance[7]	real clinical dialogue-note pairs	972 /–	452/-[1]	802k	N

Table 1. Comparable corpora for doctor-patient note generation from conversations. The majority of datasets are proprietary and unshare-able for community evaluation. (src-len = source/transcript length, target-len = target/note length, - = unreported). [1]The authors model sections of the note differently. The number of sources and note sections are different. Here we approximate the average note length by adding the average section lengths together. Average source length was approximated by averaging the sources for different sections.

or all of the note creation process, which has been linked with improved productivity, increased revenue, and improved patient-clinician interactions[4]. However such systems are both costly and, more importantly, often require a substantial investment in time from the providers in managing and training their scribes[5] – a problem that is often multiplied by the high attrition rates in the field.

One promising solution is the use of automatic summarization to capture and draft notes, before being reviewed by a clinician. This technology has attracted increasing attention in the last 5 years as a result of several key factors: (1) the improvement of speech-to-text technology, (2) widespread adoption of electronic medical records in the United States, and (3) the rise of transformer models. Several works have adopted early technology in this area, including use of statistical machine translation methods, use of RNNs, transformers, and pre-trained transformer models[6–11].

However, a massive bottleneck in understanding the state-of-the-art is the lack of publicly share-able data to train and evaluate[12]. This challenge is inherent in the required data's characteristics as (1) meeting audio and transcripts from medical encounters are not typically recorded and saved, and (2) medical information is highly personal and sensitive data and cannot be easily, ethically shared publicly. Private companies may construct or acquire their own private datasets; however, results and algorithms cannot be systematically compared. Recent ground-breaking performances by large language models such as ChatGPT and GPT4 provide promising general model solutions; however without common datasets that may be studied publicly, it would be impossible for the scientific community to understand strength, weaknesses, and future directions.

In this paper, we present the Ambient Clinical Intelligence Benchmark (ACI-BENCH) corpus. The corpus, created from domain experts, is designed to model three variations of model-assisted clinical note generation from doctor-patient conversations. These include conversations with (a) calls to a virtual assistant (e.g. required use of wake words or prefabricated, canned phrases), (b) unconstrained directions or discussions with a scribe, and (c) natural conversations between a doctor and patient. We also provide data to experiment between using human transcription and automatic speech recognition (ASR); or between ASR and corrected ASR. Table 1 shows a comparison of the 8 corpora described in state-of-the-art work. Only two other similar corpora are publicly available. MTS-dialog[13] contains ~1700 samples however its focus is on on dialogue snippets rather than full encounters. primock57[14] contains a small set of 57 full-length encounters. To our knowledge, ACI-BENCH is the largest and most comprehensive corpus publicly available for model-assisted clinical note generation.

In the following sections, we provide details of the ACI-BENCH Corpus. We (1) discuss the dataset construction and cleaning, (2) provide statistics and the corpus structure, (3) describe our content validation methods and comparison with real data, (4) quantify several diverse baseline summarization methods on this corpus.

Methods

Data creation. Clinical notes may be written by the physician themselves or in conjunction with a medical scribe or assistant; alternatively physicians may choose to dictate the contents of an entire note to a human transcriptionists or an automatic dictation tool. In cases with human intervention, scribe-assisted or transcriptionist-assisted cases, physician speech may include a mixture of commands (e.g. "newline", "add my acne template"), free-text requiring almost word-for-word copying (e.g. "To date, the examinee is a 39 year-old golf course maintenance worker")[6], or free-text communication to the medical assistance (e.g. "let's use my normal template, but only keep the abnormal parts", "can you check the date and add that in?"). With trained medical scribes participating in the clinic visit, in addition to directions from the doctor, they are expected to listen in on the patient-doctor dialogue and generate clinical note text independently. To mirror this reality, the ACI-BENCH corpus consists of three subsets representing common modes of note generation from doctor-patient conversations:

Virtual assistant (virtassist). In this mode, the doctor may use explicit terms to activate a virtual assistance device (e.g. "Hey Dragon show me the diabetes labs") during the visit. This necessitates some behavioral changes on the part of the provider.

type	annotated	example
dates (none mentioned)	Y	PSA 0.6 ng/mL[, 05/25/2021]
exam	Y	[Consitutional: Well-developed, well-nourished, in no apparent distress]
	Y	Neck: [Supple without thyromegaly or lymphadenopathy.] No carotid bruits appreciable.
medication context	Y	1 tablet [by oral route] daily
medical reasoning	Y	recommended that we obtain an MRI of the right shoulder [to evaluate for a possible rotator cuff tear].
	Y	referred her to formal physical therpay [to strengthen her right shoulder]
patient acquiescence	Y	[All questions were answered.]
	Y	[The patient understands and agrees with the recommended medical treatment plan]
review of system	Y	Ears, Nose, Mouth and Throat: [Denies ear pain, hearing loss, or discharge.] Endorses nasal congestion from allergies.
vitals	Y	[Blood Pressure:124/82 mmHg]
dates (year kept if only month mentioned)	N	03/[2022]
higher granularity problem/test/treatment	N	diabetes [type II]
	N	[3 views] of the shoulder
	N	MRI [of the head]
measurements	N	25 [mg/dl]
patient name/age	N	[John Smith] is a 53-year-old male
other names	N	he was seen by Jane [Smith, PA-C]

Table 2. Examples of unsupported text (demarked by square brackets). In the original demo data, these items were added for realism without basis in the source text, the doctor-patient conversation. Some unsupported items were purposely left unmarked in cases where removal would lead to note quality/meaning degradation. After human annotated text-span level identification, these were automatically removed from the clinical note.

Virtual scribe (virtscribe). In this mode, the doctor may expect a separate scribe entity (automated or otherwise) to help create the clinical note. This subset is characterized by pre-ambles (e.g. short patient descriptions prior to a visit) and after-visit dictations (e.g. used to specify non-verbal parts of the visit such as the physical exam or to dictate the assessment and plan). The rest of the doctor-patient conversation will be natural and undisturbed.

Ambient clinical intelligence (aci). This data is characterized by natural conversation between a patient and a doctor; without explicit calls to a virtual assistant or additional language addressed to a scribe.

Transcripts from subsets **virtassist** and **virtscribe** were created by a team of 5+ medical experts including medical doctors, physician assistance, medical scribes, and clinical informaticians based on experience and studying real encounters. Subset **aci** was created with a certified doctor and a volunteer lay person, who must role-play a real doctor-patient encounter, given a list of symptom prompts. Clinical notes were generated using an automatic note generation system and checked and re-written by domain experts (e.g. medical scribes, or physicians). The **virtscribe** dataset includes the human transcription as well as an ASR transcript; meanwhile the **virtassist** and **aci** subsets were created with only a human transcription and ASR transcript available, respectively.

Data cleaning and annotation. Our final dataset was distilled from encounters originally created for marketing demonstration purposes. During this initial dataset creation, imaginary EHR injections were placed within the note to contribute to realism, though many without basis from the conversation. Although EHR inputs, independent from data intake from a conversation, are a critical aspect of real clinical notes, in this dataset we do not model EHR input or output linkages with the clinical note (e.g. smart links to structured data such as vitals values, structured survey data, order codes, and diagnosis codes).

In order to identify unsupported information of note text to the transcript, we created systematic annotation guidelines for labeling unsupported note sentences. These unsupported information included items such as reasoning for treatment (which may not be part of the original conversation) or could be information from imaginary EHR inputs (e.g. vitals). Examples of the different types of unsupported information are included in Table 2. We tasked four independent annotators with medical backgrounds to complete this task. The partial span overlap agreement was 0.85 F1. Marked text spans were removed during automatic processing.

Because the datasets were originally created and demonstrated for a short period, as such, these notes were created under greater time constraints and less review. To ensure quality, four annotators identified and corrected note errors, such as inconsistent values. Finally, as the ACI-BENCH dataset used ASR transcripts, there were cases where the note and the transcript information would conflict due to ASR errors. For example, "hydronephosis" in the clinical note may be wrongly automatically transcribed as "high flow nephrosis". Another example may be a names; "Castillo" may be transcribed as "kastio". As part of this annotation, we tasked annotators to identify these items and provide corrections. After annotation, the data was processed such that note errors were corrected and unsupported note sentences were removed. To study the effect of ASR errors, ASR transcripts were processed into two versions: (a) original and (b) ASR-corrected (ASR outputs corrected by humans). After automatic processing, encounters were again manually reviewed for additional misspelling and formatting issues.

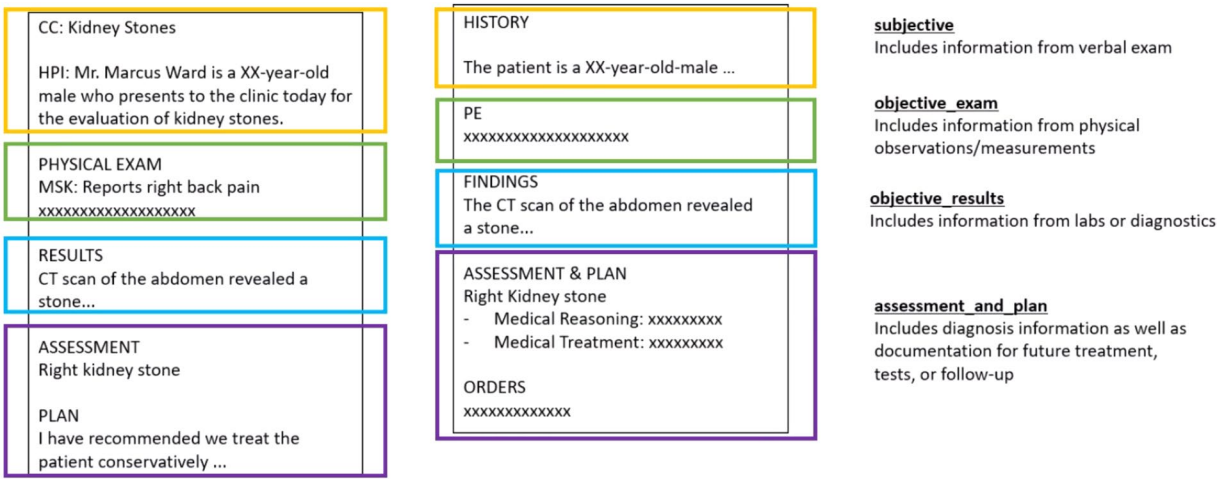

Fig. 1 Note division example. The same content in a clinical note can appear under different sections. As an example, in the left note, "past medical history" contents are written in the "history" portion of the note on the right. To separate the full note target into smaller text and minimize data sparsity problems if modeling by individual sections, notes are partitioned into separate SUBJECTIVE, OBJECTIVE_EXAM, OBJECTIVE_RESULTS, and ASSESSMENT_AND_PLAN continuous divisions. This also allows evaluation and generation at a higher granularity compared to a full note level.

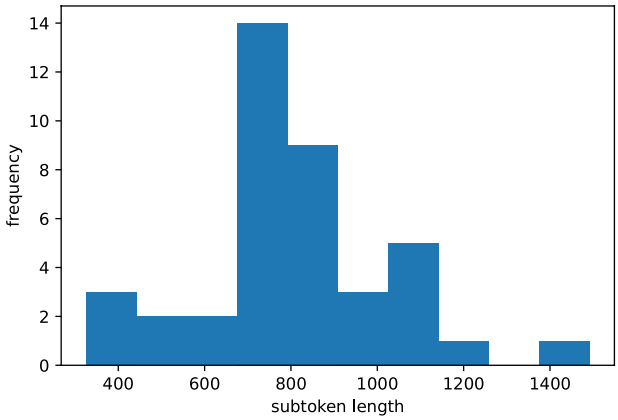

Fig. 2 BERT subtoken lengths of concatenated gold/system summaries (test1 Text-davinci-003 system) for doctor-patient dialogue to clinical note generation task. As embedding-based models require encoding the concatenated reference and hypothesis, it would be difficult to fairly evaluate the corpus using current pretrained BERT models with a 512 limit.

Note division definition. Motivated by a need to simplify clinical note structure, improve sparsity problems, and simplify evaluation, in this section, we describe our system for segmenting a full clinical note into continuous divisions.

Clinical notes are semi-structural documents with hierarchical organization. Each physician, department, and institution may have their own set of commonly used formats. However, no universal standard exists[15]. The same content can appear in multiple forms structured under different formats. This is illustrated in the subjective portions of two side-by-side notes in Fig. 1. In this example, contextual medical history appears in their own sections (e.g. "current complaint (cc)" and "history of present illness (hpi)") in the report on the left; and merged into one history section in the report on the right. These variations in structure pose challenges for both generation and evaluation. Specifically, if evaluating by fine-grained sections in the reference, it is possible that generated notes may include the same content in other sections. Likewise, generating with fine-grained sections would require sufficient samples from each section; however as not every note has every type of section – the sample size becomes sparser. Finally, it is important to note, current state-of-the-art pre-trained embedding based evaluation metrics (e.g. bertscore, bleurt, bart-score) are limited by the original trained sequence length which are typically shorter than our full document lengths. This is illustrated in Fig. 2, where for one system (Text-davinci-003) the length of the concatenated reference and system summaries will typically far exceed the typical pre-trained BERT-based 512 subtoken limit.

	train	valid	test1	test2	test3
number encounters	67	20	40	40	40
dialogue					
avg number turns	56	53	52	56	58
avg length (tok)	1301	1221	1231	1382	1334
note					
avg length (tok)	483	492	476	500	505
avg length (sentences)	48	49	47	50	50
# subjective	67	20	40	40	40
# objective_exam	64	19	40	39	39
# objective_results	53	18	32	29	27
# asessment_and_plan	67	20	40	40	40
subset					
# **virtassist**	20	5	10	10	10
# **virtscribe**	12	4	8	8	8
# **aci**	35	11	22	22	22

Table 3. Corpus statistics.

To simplify training and evaluation, as well as maintain larger samples of data, we partition notes and group multiple sections together into four divisions, as shown in Fig. 1 These divisions were inspired by the SOAP standard, where the SUBJECTIVE includes items taken during verbal exam and typically written in the chief complaint, history of present illness, and past social history; the OBJECTIVE_EXAM includes content from the physical examination on the day of the visit; the OBJECTIVE_RESULTS includes diagnostics taken prior to the visit, including laboratory or imaging results; and the ASSESSMENT_AND_PLAN includes the doctor's diagnosis and planned tests and treatments[16]. In our dataset, the divisions are contiguous and appear in the order previously introduced. Another practical benefit of partitioning the note into contiguous divisions is the greater ability to leverage pretrained sequence-to-sequence models, typically trained with shorter sequences. Furthermore, evaluation at a sub-note level allows a greater resolution for assessing performances.

Data statistics. The full dataset was split into train, validation, and three test sets. Each subset was represented in the splits through randomized stratefied sampling. The test sets 1 and 2 corresponds to the test sets from ACL ClinicalNLP MEDIQA-Chat 2023 (https://github.com/abachaa/MEDIQA-Chat-2023) TaskB and TaskC, respectively. Test 3 corresponds to TaskC of CLEF MEDIQA-SUM 2023 (https://www.imageclef.org/2023/medical/mediqa). The frequency of each data split are shown in Table 3.

Data Records

The ACI-BENCH Corpus can be found at https://doi.org/10.6084/m9.figshare.22494601[17]. Code for pre-processing, evaluation, and running baselines can be found in https://github.com/wyim/aci-bench.

Folder and naming organization. Data used in the ACL-clinicalnlp MEDIQA-CHAT and CLEF MEDIQASUM challenges are located in the *challenge_data* folder, whereas ASR experiment data is located in the *src_experiment_data* folder. Each data split has two associated files: a metadata and a data file (further described below). Train, validation, test1, test2, and test3 data files are prefixed with the following names: train, valid, clinicalnlp_taskB_test1, clinicalnlp_taskC_test2, and clef_taskC_test3, respectively. Source experiment data files offer subset-specific versions of train/validation/test in which the transcript may be the alternate forms of ASR or ASR-corrected versions. The naming convention prefix of these is according to the pattern: {split}_{subset}_{transcript-version}. Therefore, for example, train_virtscribe_humantrans.csv will give the training data from the **virtscribe** subset with the original human transcription version; whereas train_virtscribe_asr.csv will give the ASR transcription version.

Metadata files (*_metadata.csv). Metadata files include columns for the dataset name (e.g. **virtassist, virtscribe, aci**), *id*, *encounter_id*, *doctor_name*, *patient_firstname*, *patient_familyname*, *gender*, *chief complaint (cc)*, and *secondary complaints (2nd_complaints)*. Both *id* and *encounter_id* can be used to identify a unique encounter. The *encounter_id* were the identifiers used for the MEDIQA-CHAT and MEDIQASUM 2023 competitions. The *id* unique identifier will also denote a specific subset.

Transcript/Note files (*.csv). In the source-target data files, transcript and note text are given along with the dataset name and *id* or *encounter_id*. This file may be joined with the metadata files using either *id* or *encounter_id*. *encounter_id* should be used for challenge data, whereas the *id* should be used for the source experiment data.

Technical Validation

Content validation. After dataset creation and cleaning, an additional content validation step was conducted to ensure medical soundness. For each encounter, medical annotators were tasked with reviewing each symptom, test, diagnosis and treatment from the encounter. In cases where the medical annotation specialist was unsure of certain facts (e.g. can drug X be prescribed at the same time as drug Y?), the encounter undergoes two

	SUBJECTIVE	OBJECTIVE_EXAM	OBJECTIVE_RESULTS	ASSESSMENT_AND_PLAN	full
consult					
avg length (tok)	393	149	19	122	683
avg length (sentences)	35	19	2	11	66
aci-validation					
avg length (tok)	229	48	23	192	492
avg length (sentences)	24	8	4	16	49

Table 4. Data statistic comparing notes from aci-validation with a sample of real doctor-patient.

	consult	aci-corpus
dialogue		
avg length (no speaker tokens) (tok)	1505	1203
avg length (sentences)	141	80
note		
avg length (tok)	683	492
avg length (sentences)	66	49
annotation		
fraction note sentences aligned	0.84	0.95
fraction transcript sentences aligned	0.34	0.49
fraction crossing annotations	0.67	0.75
avg alignment text similarity	0.15	0.12
avg encounter dialogue-note text similarity	0.26	0.31
% note sentences with labels		
DICTATION	8	4
QA	15	43
STATEMENT	23	29
VERBALIZATION/STATEMENT2SCRIBE	17	7

Table 5. Alignment statistic comparison of aci-validation with a sample of real doctor-patient. (DICTATION: word-for-word copy-paste statements from the transcript, QA: question-answer conversation adjacency pairs, STATEMENT: conversation statements, VERBALIZATION/STATEMENT2SCRIBE: directed instructions or content to a external scribe).

possible additional reviews. Firstly, if the phenomenon in question can be searched identified from a + 3 M store of propriety clinical notes–including multiple providers across the the country (which we will refer to as the CONSULT dataset)–we deemed the information credible. Alternatively, if the information is not something that could be identified by the first approach, the question is escalated to a clinical expert annotator. Encounters with severe logical or medical problems identified by a medical annotators were removed (e.g. using a medication for urinary tract infection for upper respiratory infection).

Comparison with real data. To study differences between the ACI-BENCH dataset and a set of real encounters, we conduct statistical comparison with 163 randomly chosen family medicine clinical encounters (including pairs human transcriptions and corresponding clinical notes) with in-depth alignment annotation, from the CONSULT dataset. Tables 4, 5 show the statistical comparison between the 20 encounters in the validation set (aci-validation) and the CONSULT encounters. In general, the ACI-BENCH dataset had on average shorter notes, at 492 tokens versus 683 tokens for the consult dataset. Except for the OBJECTIVE_RESULTS division, every division was longer in the consult data (Table 4). The ACI-BENCH dataset also exhibits shorter dialogue lengths, by approximately 100 tokens and 20 sentences; as well a shorter notes by approximately 100 tokens (Table 5). One reason for the shorter note length is our removal of unsupported note text.

We additionally annotated for alignments of data between the source and target on the validation set (20 encounters) and consult set, similar to that of previous work[18]. This annotation marks associations between note sentences and their corresponding source transcript sentences. Unmarked note sentences indicate that a sentence may be purely structural (e.g. section header) or may include unsupported content. Likewise, unmarked transcript sentences may indicate that the content is superfluous. Comparing the portions of annotated alignments in separate corpora gives indications of corpora similarity with respect to relative content transfer. Other useful metrics which provide measures of alignment/generation difficulty include: (a) the fraction of alignment crossings (whether content appear monotonically versus "out-of-order"/"crossing")[19]), (b) the similarity of corresponding text segments, and (c) percentage of transcript speech modes. The results of these comparisons are shown in Table 5.

Labeled alignment annotations show that approximately the same fractions of dialogue and note sentences were labeled (0.34 and 0.49 transcript, 0.84 and 0.95 note for the consult and ACI-BENCH corpus respectively);

with a high 0.95 fraction for the ACI-BENCH corpus, as designed by the removal of unsupported text. With shorter transcripts (1203 tokens in ACI-BENCH vs 1505 tokens in the CONSULT set), the ACI-BENCH corpus also had a 15% more aligned transcript sentences. The text similarity (Jaccard unigram) of alignments were similar (0.15 and 0.12) as was the fraction of crossing annotations (0.67 and 0.95) for the CONSULT and ACI-BENCH corpus respectively; though the dialogue-note document similarity was higher in the ACI-BENCH corpus.

The percentage of note sentences annotated with different labels show across the board lower percentages in the CONSULT data. This is explainable as the transcript length and thus the percentage of note sentences annotated with a certain label will decrease. However, it is interesting to show that the ACI-BENCH corpus had a higher percentage of note sentences coming from question-answer paired transcript sentences and conversation statements rather than dictation/statement2scribe. For example while in the CONSULT dataset, important QA makes up twice as much transcript sentences as in dictation (15% and 8%), in the ACI-BENCH dataset there are ten times more QA labeled sentences than dictation (43% vs 4%). Meanwhile in the CONSULT dataset, transcript sentences identified with an alignment using the "statement" tag was about three times that of dictation, however this was about seven times in the ACI-BENCH corpus. Together, this data suggests that the ACI-BENCH corpus may be slightly less challenging in terms of documents lengths and has a skew towards question-answer and statements information content; though the magnitudes in lengths and similarity are comparable.

Baseline experiments. In this section, we present our baseline experiments designed to benchmark the ACI-BENCH Corpus. These experiments encompass various note-generation tasks and incorporate state-of-the-art note-generation techniques. To assess the robustness of note-generation techniques, we also examine the impact of different clinical doctor-patient dialogue transcript generation methods with and without human correction on the quality of automatically generated clinical notes derived from these transcripts.

Note generation models. The experiments on note-generation models to benchmark the ACI-BENCH Corpus are listed below:

Transcript-copy-and-paste Previous research finds taking the longest sentence[20] as dialogue summarization is a good baseline. In the spirit of this approach, we adopt several variations to generate the clinical note: (1) the longest speaker's turn, (2) the longest doctor's turn, (3) the first two and the last ten speaker's turns, (4) the first two and the last ten doctors turns and (5) the entire transcript.
Retrieval-based Borrowing from retrieval-based response generation[21], we pose a simple baseline that retrieves a relevant note in the training corpus rather than generating new text. To generate a clinical note for a new transcript, we employ transcript UMLS concept set similarity to retrieve the most similar transcript from the train set. The note that corresponds to this transcript in the training set is selected as the summarization for the new transcript, based on the assumption that the semantic overlap between the UMLS concepts in the two transcripts is a reliable indicator of their content similarity. Following the same manner, we adopt a similar retrieval-based method on the document embedding similarity from the spaCy English natural language process pipeline[22].
BART-based We employ the SOTA transformer model, bidirectional autoregressive transformer (BART)[23]. We also include its two variants: (1) a version with continued pre-training on PubMed abstract[24], aimed at learning domain-specific language and knowledge, and (2) a version fine-tuned on the SAMSum corpus[25], designed to enhance the model's performance on conversational summarization tasks. For all BART-based models, we use the BART-Large version. It is important to note that although BART and BioBART have the same model structure, they possess distinct tokenizers and vocabulary sizes. These differences play a significant role in determining their respective performance on the ACI-BENCH corpus. The corresponding fine-tuning parameters can be found in the Supplementary Information. BART-based models have the same limit of 1,024 tokens.
LED-based We leverage the Longformer-Encoder-Decoder (LED) architecture[26], which incorporates an attention mechanism that can scale up to longer sentences. LED-based models have the same limit of 16 K tokens. Because the transcript is long, LED overcomes the sentence length limit from BART. We also include its variant, which is finetuned on the Pubmed dataset[27], to enhance the model's summarization ability in the biomedical context. The corresponding fine-tuning parameters can be found in the Supplementary Information.
OpenAI models We experimented with the latest OpenAI models and APIs (https://platform.openai.com/docs/models): (i) Text-davinci-002, (ii) Text-davinci-003, (iii) ChatGPT (gpt-3.5-turbo), and (iv) GPT-4. The first three models have the same limit of 4,097 tokens, shared between the prompt and the output/summary, whereas GPT-4 allows 32k tokens. We used the following prompt:

- Prompt: "summarize the conversation to generate a clinical note with four sections: HISTORY OF PRESENT ILLNESS, PHYSICAL EXAM, RESULTS, ASSESSMENT AND PLAN. The conversation is:"

To allow adequate division detection, we added some light rule-based post-processing, adding endlines before and after for each section header. This post-processing described in the Supplementary Information.

Full-note- vs division-based note-generation approaches. In the cases of the fine-tuned pre-trained models, we investigated note generation with two overall approaches: full note generation versus division-based generation and concatenation. The first approach generates a complete note from the transcript using a single model or approach. The latter approach is motivated by the long input and output lengths of our data – which may exceed that of those pre-trained models are typically trained for. To this end, full notes were divided into the SUBJECTIVE, OBJECTIVE_EXAM, OBJECTIVE_RESULTS, and ASSESSMENT_AND_PLAN divisions using a rule-based regular-expression section detection. As the notes were followed a handful of regular patterns, this section detection was highly performant. In cases where certain sections were missing, an *EMPTY* flag was used as the

Model	ROUGE-1	ROUGE-2	ROUGE-L	MEDCON
Transcript-copy-and-paste				
longest spearker turn	27.84	9.32	23.44	32.37
longest doctor turn	27.47	9.23	23.20	32.33
12 speaker turns	33.16	10.60	30.01	39.68
12 doctor turns	35.88	12.44	32.72	47.79
transcript	32.84	12.53	30.61	55.65
Retrieval-based				
$train_{UMLS}$	43.87	17.55	40.47	33.30
$train_{sent}$	41.59	15.50	38.20	26.17
BART-based				
BART	41.76	19.20	34.70	43.38
BART (Division)	51.56	24.06	45.92	47.23
BART + FT_{SAMSum}	40.87	18.96	34.60	41.55
BART + FT_{SAMSum} (Division)	**53.46**	**25.08**	**48.62**	48.23
BioBART	39.09	17.24	33.19	42.82
BioBART (Division)	49.53	22.47	44.92	43.06
LED-based				
LED	28.37	5.52	22.78	30.44
LED (Division)	34.15	8.01	29.80	32.67
LED + FT_{pubMed}	27.19	5.30	21.80	27.44
LED + FT_{pubMed} (Division)	30.46	6.93	26.66	32.34
OpenAI (wo FT)				
Text-Davinci-002	41.08	17.27	37.46	47.39
Text-Davinci-003	47.07	22.08	43.11	57.16
ChatGPT	47.44	19.01	42.47	55.84
GPT-4	51.76	22.58	45.97	**57.78**

Table 6. Results of the summarization models evaluated at the full note level, test set 1. Simple retrieval-based methods provided strong baselines wih better out-of-the-box performances than LED models and full-note BART models. In general for BART and LED fine-tuned models, division-based generation worked better. OpenAI models with simple prompts were shown to give competitive outputs despite no additional fine-tuning or dynamic prompting.

output. Each division generation model was separately fine-tuned. The final note was created by concatenating the divisions.

Automatic evaluation metrics. We employ a variety of widely-used automatic evaluation metrics to evaluate performances in different perspectives. Specifically, we measure at least one lexical n-gram metric, an embedding-based similarity metric, a learned metric, and finally an information extraction metric. We evaluate the note generation performance both in the full note and in each division.

For the ngram-based lexical metric, we compute ROUGE[28](1/2/-L), which computes unigram, bigram, and the longest common subsequence matches between reference and candidate clinical notes. For an embedding-based metric, we applied BERTScore[29] which greedily matches contextual token embeddings from pairwise cosine similarity. BERTScore efficiently captures synonym and context information. For a model-based learned metric, we used BLEURT[30], which is trained for scoring candidate-reference similarity. Additionally, we incorporate a medical concept- based evaluation metric (**MEDCON**) to gauge the accuracy and consistency of clinical concepts. This metric calculates the F1-score to determine the similarity between the Unified Medical Language System (UMLS) concept sets in both candidate and reference clinical notes. This is similar to the CheXpert evaluation for radiology summarization however our concepts are not restricted to 14 predetermined categories, and do not include weightings or assertion status[31]. The extraction of UMLS concepts within clinical notes is performed using a string match algorithm applied to the UMLS concept database through the QuickUMLS package[32]. To ensure clinical relevance, we restrict the **MEDCON** metric to specific UMLS semantic groups, designated as *Anatomy*, *Chemicals &Drugs*, *Device*, *Disorders*, *Genes & Molecular Sequences*, *Phenomena* and *Physiology*. To consolidate the various evaluation metrics, we first take the average of the three ROUGE submetrics as ROUGE, and the average of ROUGE, BERTScore, BLEURT, and **MEDCON** scores as the final evaluation score. Because BERTScore and BLEURT are limited by their pre-trained embedding length, we only use these evaluations for the division-based evaluation.

Results. We fine-tune the models on the train set and select the best trained model based on evaluation on the validation set. Performances were evaluated on three test sets. Test sets 1 and 2 correspond to the test sets from ACL ClinicalNLP MEDIQA-Chat 2023 TaskB full-note generation and TaskC dialogue generation, respectively. Test 3 corresponds to CLEF MEDIQA-SUM 2023 Subtask C full-note generation. Our test 1 full note evaluation results can be found in Table 6. Per-division SUBJECTIVE, OBJECTIVE_EXAM, OBJECTIVE_RESULTS, AND

Model	Evaluation score on the subjective division						
	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	BLEURT	MEDCON	Average
Retrieval-based							
$train_{UMLS}$	41.70	23.45	31.64	72.10	39.01	23.04	41.60
$train_{sent}$	41.12	20.62	29.20	70.78	37.94	18.86	39.47
BART-based							
BART	48.19	25.81	30.13	68.93	43.83	44.41	47.97
BART (Division)	47.25	26.05	31.21	70.05	43.55	44.20	48.16
BART + FT_{SAMSum}	46.33	25.52	29.88	68.68	**45.01**	43.21	47.70
BART + FT_{SAMSum} (Division)	**52.44**	**30.44**	**35.83**	**72.41**	44.51	**47.84**	**51.08**
BioBART	45.79	23.65	28.96	68.49	41.09	41.10	45.87
BioBART (Division)	46.29	25.99	32.43	70.30	42.99	41.14	47.33
LED-based							
LED	24.81	5.29	11.00	55.60	30.68	20.19	30.04
LED (Division)	31.27	8.31	15.99	56.94	25.40	24.03	31.22
LED + FT_{pubMed}	23.48	4.72	10.49	54.46	20.32	17.91	26.40
LED + FT_{pubMed} (Division)	26.03	6.17	12.93	56.41	19.19	20.46	27.78
OpenAI (wo FT)							
Text-Davinci-002	29.73	12.38	20.13	58.98	36.70	32.47	37.22
Text-Davinci-003	33.29	15.24	23.76	60.63	38.06	36.14	39.73
ChatGPT	32.70	14.05	22.69	65.14	39.48	38.21	41.49
GPT-4	41.20	19.02	26.56	63.34	43.18	44.25	44.93

Table 7. Results of the summarization models on the subjective division, test set 1. BART-based models generated at both full note and division levels had similar levels of performances, which were in general better than the other model classes. As in the full note evaluation, retrieval-based methods provided competitive baselines.

ASSESSMENT_AND_PLAN results for test 1 are accounted for in Tables 7–10. In the main body of this paper, we discuss the results of test 1 which was used for our first full note generation task challenge (TaskB in https://github.com/abachaa/MEDIQA-Chat-2023). We will first provide an overview of the model performance in both full-note and division-based evaluations. We will then describe each model type's performance. For reference, we provide the results of test 2 and test 3 in the Supplementary Information.

In the full-note evaluation, the BART + FT_{SAMSum} (Division) model achieved the highest ROUGE scores, with 53.46 for ROUGE-1, 25.08 for ROUGE-2 and 48.62 for ROUGE-L. This is because when BART + FT_{SAMSum} (Division) model was fine-tuned on our ACI-BENCH training set, it learned more specific clinical jargon in the ACI-BENCH corpus, such as accurate subsection headers ("CHIEF COMPLAINT", "HISTORY OF PRESENT ILLNESS",…) and physical examination results ("- Monitoring of the heart: No murmurs, gallops…"). On the contrary, GPT-4 demonstrated the highest **MEDCON** evaluation score of 57.78, while achieving the second to third-best performance in ROUGE scores, with 51.76 for ROUGE-1, 22.58 for ROUGE-2 and 45.97 for ROUGE-L. The great performance can be attributed to the model's gigantic size, intensive pretraining, huge context size, and great versatility. GPT-4 captured many relevant clinical facts and thus had the highest **MEDCON**. However, since it was not specifically fine-tuned for the ACI-BENCH corpus clinical note format, it exhibited slightly inferior performance in capturing the structured ACI-BENCH clinical notes. An example of a note generated from different models can be found in the Supplementary Information. Interestingly, the retrieval-based baselines showed very competitive ROUGE performances out-of-the-box with ROUGE-L of 40.47 F1 for and 38.20 F1 for the UMLS and sentence versions respectively. Furthermore, the simple transcript copy-and-paste baselines produced high starting points that out-performed untreated LED-based models. For example, simply copying the transcript achieved a 30.61 F1 ROUGE-L and 55.65 F1 **MEDCON** score, whereas the division based LED model achieved 29.80 F1 and 32.67 F1.

In division-based evaluations, we found that different models achieved the highest average score across different note divisions, BART + FT_{SAMSum} (Division) scored 51.08 in the SUBJECTIVE division (Table 7), Text-davinci-003 reached 55.30, 48.90 and 46.19 in the OBJECTIVE_EXAM (Table 8), OBJECTIVE_RESULTS (Table 9), and ASSESSMENT_AND_PLAN (Table 10) divisions, respectively. These results indicate that all three models can be good candidates for the note-generation task. However, since BART + FT_{SAMSum} (Division) required fine-tuning and Text-davinci-003 did not, the latter two models demonstrated greater potential. A few additional examples for Text-davinci-003 could potentially enhance their performance, by enabling the models to learn specific clinical jargon in each division.

In comparing the full-note and division-based note-generation approaches, our experiments demonstrated that, for our pretrained BART- and LED-based models, division-based note-generation methods resulted in significant improvements over full-note-generation methods. These improvements ranged from 1 to 14 point increases in both ROUGE and MEDCON evaluations for the full-note-based evaluation. This finding implies that breaking down a complex summarization problem into smaller divisions effectively captures more critical information. For division-based evaluations, the increase is not obvious for the SUBJECTIVE divisions, but

Model	Evaluation score on the objective_exam division						
	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	BLEURT	MEDCON	Average
Retrieval-based							
$train_{UMLS}$	43.43	25.77	36.74	**74.63**	40.96	24.63	43.88
$train_{sent}$	37.02	19.58	31.20	71.47	35.83	14.52	37.77
BART-based							
BART	0.56	0.36	0.56	40.25	10.66	0.00	12.85
BART (Division)	49.77	31.63	38.92	73.75	44.19	34.80	48.21
BART + FT_{SAMSum}	6.22	3.74	5.21	44.33	14.82	4.14	17.09
BART + FT_{SAMSum} (Division)	47.73	29.51	36.98	73.41	42.86	35.91	47.56
BioBART	2.57	1.04	1.68	42.10	12.38	1.22	14.36
BioBART (Division)	42.51	26.15	32.19	71.57	42.18	29.55	44.23
LED-based							
LED	0.00	0.00	0.00	0.00	14.87	0.00	3.72
LED (Division)	27.03	7.96	16.88	54.48	14.47	18.84	26.27
LED + FT_{pubMed}	0.00	0.00	0.00	0.00	14.87	0.00	3.72
LED + FT_{pubMed} (Division)	20.24	6.30	12.14	54.13	12.67	18.07	24.44
OpenAI (wo FT)							
Text-Davinci-002	43.68	22.31	35.03	68.25	45.68	35.41	45.75
Text-Davinci-003	**54.17**	**32.42**	**44.54**	73.40	**51.29**	**52.79**	**55.30**
ChatGPT	49.44	27.29	38.60	71.39	49.39	48.95	52.04
GPT-4	50.11	28.20	40.43	71.79	51.11	42.59	51.27

Table 8. Results of the summarization models on the OBJECTIVE_EXAM division, test set 1. BART and LED full note generation models suffered a significant drop at the OBJECTIVE_EXAM. This may be attributable to the lower amounts of content required to be generated, the appearance of text later in the sequence, as well as the higher variety of structures. The OpenAI were in general better performant with BART division-based models as next best.

Model	Evaluation score on the objective_results division						
	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	BLEURT	MEDCON	Average
Retrieval-based							
$train_{UMLS}$	30.26	14.89	29.87	66.24	37.25	8.91	34.35
$train_{sent}$	40.52	18.21	38.87	**73.33**	45.79	12.45	41.03
BART-based							
BART	0.00	0.00	0.00	0.00	5.45	0.00	1.36
BART (Division)	30.48	19.16	27.80	66.64	43.07	21.56	39.27
BART + FT_{SAMSum}	20.79	0.46	20.67	54.54	28.32	0.77	24.40
BART + FT_{SAMSum} (Division)	29.45	18.01	26.63	66.43	40.75	20.17	38.01
BioBART	17.50	0.00	17.50	52.44	25.33	0.00	22.36
BioBART (Division)	35.38	14.33	32.79	68.40	47.63	15.69	39.81
LED-based							
LED	0.00	0.00	0.00	0.00	5.45	0.00	1.36
LED (Division)	14.04	4.97	11.08	48.86	9.61	7.86	19.09
LED + FT_{pubMed}	0.00	0.00	0.00	0.00	5.45	0.00	1.36
LED + FT_{pubMed} (Division)	10.48	3.64	8.32	42.43	7.13	8.86	16.48
OpenAI (wo FT)							
Text-Davinci-002	41.48	20.12	39.95	70.61	50.79	24.42	44.92
Text-Davinci-003	**44.92**	**25.21**	**43.84**	72.35	**55.87**	**29.37**	**48.90**
ChatGPT	34.50	17.75	30.84	66.68	48.51	22.28	41.29
GPT-4	37.65	19.94	35.73	68.33	48.50	26.73	43.67

Table 9. Results of the summarization models on the OBJECTIVE_RESULTS division, test set 1. Similar to OBJECTIVE_EXAM, BART and LED full note generation models suffered a significant drop at the OBJECTIVE_RESULTS division. This may be attributable to the higher sparsity of this division, low amounts of content (sometimes only 2-3 sentences), and the appearance of text later in the sequence. The OpenAI were in general better performant with BART division-based models as next best.

Model	Evaluation score on the assessment_and_plan division						
	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	BLEURT	MEDCON	Average
Retrieval-based							
$train_{UMLS}$	**44.59**	**21.50**	**29.66**	**70.39**	44.77	24.70	42.94
$train_{sent}$	41.28	19.73	28.02	69.48	43.18	18.79	40.28
BART-based							
BART	0.00	0.00	0.00	0.00	29.05	0.00	7.26
BART (Division)	43.31	20.59	26.55	67.49	40.99	32.30	42.73
BART + FT_{SAMSum}	1.52	0.49	0.87	35.38	19.79	1.00	14.28
BART + FT_{SAMSum} (Division)	43.89	21.37	27.56	68.09	41.96	31.33	43.08
BioBART	0.00	0.00	0.00	0.00	29.05	0.00	7.26
BioBART (Division)	42.44	19.44	26.42	67.57	43.88	31.12	43.00
LED-based							
LED	0.00	0.00	0.00	0.00	29.05	0.00	7.26
LED (Division)	28.23	6.13	12.44	55.75	27.78	21.94	30.27
LED + FT_{pubMed}	0.00	0.00	0.00	0.00	29.05	0.00	7.26
LED + FT_{pubMed} (Division)	28.00	5.99	13.07	55.68	20.95	25.01	29.33
OpenAI (wo FT)							
Text-Davinci-002	30.90	12.27	21.44	61.01	44.98	35.04	40.64
Text-Davinci-003	35.41	14.86	25.38	63.97	49.18	**46.40**	**46.19**
ChatGPT	36.43	12.50	23.32	63.56	48.21	43.71	44.89
GPT-4	38.16	14.12	24.90	64.26	**49.41**	42.36	45.44

Table 10. Results of the summarization models on the ASSESSMENT_AND_PLAN division, test set 1. Similar to OBJECTIVE_EXAM and OBJECTIVE_RESULTS, BART and LED full note generation models suffered a significant drop at the OBJECTIVE_RESULTS division. This may be attributable to the appearance of text later in the sequence. The OpenAI were in general better performant with BART division-based models as next best.

around 20 percent in the average score for OBJECTIVE_EXAM, OBJECTIVE_RESULTS and ASSESSMENT_AND_PLAN divisions. This can be attributed to the generation of the latter three divisions at the end of clinical notes, which often exceeds the word length of typical summarization tasks that BART-based and LED-based models are used for. Additionally, since some notes in the training set lack these divisions, the note-generation models struggle to learn the division structure during fine-tuning from the full note. As the division of clinical notes is identified by a rule-based division header extraction method, even when the information from a specific division is generated as a few sentences, the corresponding division information cannot be detected by the evaluation program.

Our observations on the performance for each type of model are summarized below:

Transcript-copy-and-paste models are only evaluated in the full note. It demonstrated suboptimal performance, which is around 17 points less than the best ROUGE scores. This is primarily because transcripts from doctor-patient dialogues serve to facilitate doctor-patient interactions with questions, answers, and explanations related to various health phenomena. In contrast, clinical notes, which are created by and intended for healthcare professionals, generally follow the SOAP format to convey the information concisely and accurately. Therefore, transcripts and notes can differ significantly in terms of terminology, degree of formality, relevance to clinical issues, and the organization of clinical concepts. On the other hand, the original transcript often achieves the fourth highest score in **MEDCON** evaluation at 55.65, owing to its ability to capture relevant UMLS concepts explicitly mentioned within the transcript.

Retrieval-based models have the best BERTScore in OBJECTIVE_EXAM, OBJECTIVE_RESULTS AND ASSESSMENT_AND_PLAN divisions in test 1, with around 1 to 5 points increase over the best BART-based and OpenAI models. They also have shown sometimes promising results with the second and first average scores in OBJECTIVE_RESULTS AND ASSESSMENT_AND_PLAN divisions from test 3. This is because clinical notes with similar transcripts tend to have more similar clinical notes, especially when OBJECTIVE_RESULTS sections use standard phrasing and templates, or in scenarios where patients share common symptoms and health examinations from different medical problems. However, their performance in the **MEDCON** evaluation metric is often poor because of the less accurate patient-specific medical conditions. As a result, these models may perform well in non-**MEDCON** evaluation metrics but may not produce accurate **MEDCON** evaluations.

BART-based models demonstrated superior performance. In full-note evaluation, BART + FT_{SAMSum} (Division) had the best ROUGE score performance with **MEDCON** evaluation scores only secondary to OpenAI models. In SUBJECTIVE division, BART + FT_{SAMSum} (Division) had top performance in all scores except BLEURT. These findings suggest that using a model fine-tuned on a similar dataset serves as a solid foundation for summarization tasks. Meanwhile, BioBART exhibits a comparatively weaker performance than BART, which could be attributed to the choice of vocabularies, tokenizers, and consequently, the quality of contextual embeddings. For BART-based models, the division-based note-generation approach improved the performance from the full-note-generation approach with around a 5 to 40 points increase in all division-based average scores. This implies that dividing the complex note-generation tasks into simpler subtasks can boost model performance.

Test set	Bart Fine-tuning	Test Split	ROUGE-1	ROUGE-2	ROUGE-L	MEDCON
1	train	ASR	48.61	18.94	41.74	42.63
	$+\text{train}_{\text{ASR}}$	ASR	**49.70**	19.96	43.82	41.96
	train	human	48.28	**20.09**	**43.98**	**46.13**
	$+\text{train}_{\text{ASR}}$	human	48.50	19.52	43.59	42.85
2	train	ASR	**51.29**	**21.31**	43.76	**45.21**
	$+\text{train}_{\text{ASR}}$	ASR	50.42	21.30	**44.68**	43.71
	train	human	50.11	20.80	44.44	43.35
	$+\text{train}_{\text{ASR}}$	human	48.44	20.47	43.68	44.28
3	train	ASR	50.41	**20.01**	43.79	**49.91**
	$+\text{train}_{\text{ASR}}$	ASR	49.22	19.72	43.19	44.18
	train	human	**50.86**	19.50	**44.59**	45.48
	$+\text{train}_{\text{ASR}}$	human	47.42	18.42	42.67	44.72

Table 11. Model performance on different test sets splits, comparison between *virtscribe* dialogues with ASR and human transcript. The model finetuned on the train set is the BART + $\text{FT}_{\text{SAMSum}}$ (Division) fine-tuned with 10 epochs on the original train set, as in the baseline methods. The train + $\text{train}_{\text{ASR}}$ model refers to the BART + $\text{FT}_{\text{SAMSum}}$ (Division) finetuned for 3 more epochs on the *virtscribe* with ASR split of the train set.

LED-based models were generally inferior to that of BART-based models with around 15 to 40 points lower scores in full-note ROUGE and MEDCON scores. We observed that compared with the BART-based models, the LED-based models generate notes with worse fluency, less essential clinical information, and poorer division structure. On the other hand, the effect of the division method on LED-based models was similar to that on BART-based models, which lead to a 1 to 9 points increase in full-note ROUGE and MEDCON scores and a 2 to 25 points increase in division-based average scores.
OpenAI models exhibited good general performance and using a generic prompt, without fine-tuning. GPT4 outperformed other OpenAI models, at around 10 ROUGE-1 F1 points in full-note evaluation. This is consistent as GPT4 is known to have been trained with more parameters and has had shown to have made impressive performances across a variety of human tasks[33,34]. While Text-davinci-003 and ChatGPT were within 4 ROUGE-1 points in test 1, there were larger 4–9 point gaps in test 2 and 3 respectively. This information combined with the relatively stable ROUGE-1 score for GPT4 (at around ~50 Rouge-1), suggests that the earlier models had more unstable performances. Assessing the division-based performances, we see the relative ranking of the OpenAI's were more variable (with the exception of Text-davinci-002 consistently performing below the other models).

Effect of ASR vs human transcription and correction. In practice, automatic speech recognition (ASR) is widely deployed because it provides an affordable, real-time text-based transcript. However, the quality of ASR is usually worse than the human transcript, influenced by its model type, hardware, and training corpus. To study the effect of ASR vs human transcription on clinical note generation from dialogue, we evaluate the note-general model performance on transcripts generated from these two approaches. We compare the performance between human transcription versus ASR for the **virtscribe** subset of the data; and ASR versus ASR-corrected in the **aci** subet. We conduct this ablation study with one of the best models in the previous section, BART + $\text{FT}_{\text{SAMSum}}$ (Division), and compare the result on the split of the three test sets. To study the difference with human transcription versus ASR for the **virtscribe** subset, we experiment with feeding the raw ASR transcripts instead of it's original human transcription. We also fine-tune the model further to adapt to the ASR version by additionally learning for an additional 3 epochs with the same parameters using the ASR version of the **virtscribe** train set. To understand effects of the train/decode discrepancies, we evaluate the results of feeding in the original human-transcription source as well as ASR versions to both the original fine-tuned model and the further ASR-fined tuned model. The results of the **virtscribe** source experiments are presented in Table 11. We observed that the model setup with best ROUGE-1/2/L and MEDCON scores are different for each test set. Namely, BART + $\text{FT}_{\text{SAMSum}}$ (Division) with transcripts generated by ASR from **virtscribe** dialogues do not exhibit outstanding differences with the human transcription (when using ASR source input performance dropped to 41.74 F1 ROUGE-L instead of 43.98 ROUGLE-L with the original model human transcript input for test1). Further fine-tuning BART + $\text{FT}_{\text{SAMSum}}$ (Division) with ASR notes in the train set also did not greatly improve the performance (fine-tuning improved 2 points in F1 to 43.82 F1 for an ASR transcript source, and a minimal drop to 43.59 when applying the original human transcription source). This indicates that ASR and the human transcript do not have a remarkable impact on the note-generation performance from dialogue with **virtscribe**.

To study the effect of ASR versus ASR-corrected, we conduct similar experiments for the **aci** subset by substituting the original ASR transcripts with corrected versions. The results of these experiments are shown in Table 12, we also observed that the model setup with best ROUGE-1/2/L and **medcon** scores are different for each test set. The ASR-corrected did not exhibit more outstanding improvement from the original ASR on the BART + $\text{FT}_{\text{SAMSum}}$ (Division)'s note generation performance (with approximately 1 F1 point difference amongst all the test sets and evaluation versions and metrics). Further fine-tuning BART + $\text{FT}_{\text{SAMSum}}$ (Division) with ASRcorr notes in the train set also did not substantially change performance. This indicates that those ASR errors corrected by humans do not have a remarkable impact on the note generation performance.

Test set	Bart Fine-tuning	Test Split	ROUGE-1	ROUGE-2	ROUGE-L	MEDCON
1	train	ASR	54.03	24.19	48.09	45.97
	$+\text{train}_{\text{ASRcorr}}$	ASR	**54.10**	24.47	47.92	47.17
	train	ASRcorr	54.04	24.35	**48.19**	46.00
	$+\text{train}_{\text{ASRcorr}}$	ASRcorr	54.04	**24.50**	47.85	**47.36**
2	train	ASR	51.62	23.05	46.01	45.31
	$+\text{train}_{\text{ASRcorr}}$	ASR	53.14	24.47	47.43	45.34
	train	ASRcorr	51.56	23.13	45.97	**45.98**
	$+\text{train}_{\text{ASRcorr}}$	ASRcorr	**53.44**	**24.51**	**47.66**	45.47
3	train	ASR	53.07	**23.53**	**47.73**	45.27
	$+\text{train}_{\text{ASRcorr}}$	ASR	52.53	23.14	47.06	**46.38**
	train	ASRcorr	**53.13**	23.48	47.63	45.52
	$+\text{train}_{\text{ASRcorr}}$	ASRcorr	52.38	23.00	46.87	45.78

Table 12. Model performance on different test sets splits, comparison between *aci* dialogues with ASR and ASRcorr transcript. The model finetuned on the train set is the BART + $\text{FT}_{\text{SAMSum}}$ (Division) fine-tuned with 10 epochs on the original train set, as in the baseline methods. The train + $\text{train}_{\text{ASRcorr}}$ model refers to the BART + $\text{FT}_{\text{SAMSum}}$ (Division) finetuned for 3 more epochs on the *aci* with ASRcorr split of the train set.

In summary, our investigation of ASR versus human transcription shows that although ASR can generate errors in the transcript, those errors do not have a remarkable impact on the note-generation performance and are thus tolerable by our current model setting. However, this could be due to our automatic evaluation metrics evaluating the n-grams and clinical facts with uniform weights. In clinical practice, some particular medical fact errors from the ASR can have a non-trivial impact.

Usage Notes

Limitations and challenges. There are several limitations to this work. The data here is small and produced synthetically by medical annotators or patient actors in a single institution. Therefore, this dataset may not cover in a statistically representative way, all health topics, speech variations, and note format variations present in the real world. We describe below the specific context and challenges of these representation issues and how they may be mitigated.

Dialogue patterns and health topic variations due to limited underlying population. While the subjects in this dataset are fictitious persons, the creation of this dataset was accomplished by a limited number of content creators which influences the variations of linguistic dialogue patterns and health topics created. Specifically, the collections were conducted in a Nuance product demonstration studio by Nuance employees in Eastern Massachusetts. Thus, personas may have varying degrees of dialect and speech patterns consistent with the US Northeast. As Nuance hires employees from diverse backgrounds, some speakers may incidentally have linguistic patterns or accents consistent with their place of origin. In fact, in linguistics, it is recognized that speech cues including lexical, syntactic, pragmatic (e.g. what a speaker implies) and paralinguistic (e.g. pitch tone), can act as a social identity marker for an individual, which is an accumulation of the persons' experiences including age, gender, class, social identity, etc[35,36]. Such patterns of information exchange and dialogue may also affect the length of a conversation. Thus, based on either the doctor or patient, the same medical concern may be discussed more or less extensively (leading to different transcript lengths), using different turn-taking and dialogue acts, and employing particular expressions and vocabularies. One limitation in our original works was that content creators' age, academic achievement, place of origin, and areas/durations of schooling and residence were not captured. These and other factors may contribute to speech and language patterns[37]. On the other hand, social-economic status has been shown to be a factor on the types of health complaints of patient populations[38,39].

Purposefully collecting such details and modeling personas to fit different identified characteristics would be a meaningful improvement in a next iteration of this type of dataset. That said, this incorporation is non-trivial. If such a dataset was constructed as a persona content creation task, annotators may project their own ideas of represented populations which may add bias. Furthermore, no absolute deterministic relationship exists between an identified personal characteristics and speech patterns. For example, although a patient or persona may have been raised and educated outside the United States, the person's English understanding, speech, and writing may be close to that of a Californian American, depending on factors such as the age of immigration, years of formal schooling, subjects' own aptitude for recognizing pronunciations and interest in social activities[40]. Thus, even given this information, statistical significance can only be achieved if the dataset is at a large enough scale. At-scale production of synthetic encounters is difficult as such a dataset would require a large number of diverse highly qualified trained personnel, such as doctors (whose priority is their field of training, medicine). On the other hand, collecting, storing, and sharing a large collection of real patient data is a highly legally and ethically challenging endeavor.

With additional sub-population information, automatic note generation can be tuned and evaluated for different socio-economic categories to understand disparities in generation or evaluation performance.

Clinical note variations. In this dataset, there were only a hand full of variations in note format. For example, some notes contained chief complaints, but others did not; certain notes include longer history of present illness

sections, while others embed symptom context in the ASSESSMENT_AND_PLAN. The small number of note format variations was intentional for practical purposes: consistent training and demonstration of clinical note generation.

However, though this note structure regularity makes the task tractable, especially in a limited dataset, this phenomenon does not reflect reality, as even different providers practicing in the same organization and specialty may have vastly different note structure and content setup.

To address this problem, one area of expansion is to create multiple references. In fact, one of the highest standards of human evaluation on summarization, the pyramid[41] annotation scheme, requires multiple reference summaries. In this specific case, expansion can be in two settings: (a) content variation with the same note structure and (b) content variation with different note structures. These multiple references can then be used for investigating evaluation metrics (especially for long documents and in this special case of semi-structured setting) and well as be used for developing natural language generation algorithms which can control and evaluate for generation style[42].

EMR data incorporation. Clinical notes are generated often in conjunction with structured information from the electronic medical record (EMR). This may include chief complaints as registered during the appointment creation, vitals fed directly from instruments (e.g. blood pressure), templates specific to a doctor or practice pre-populated prior or during note entry, or structured links to diagnoses and orders for referrals, tests, and treatments. Interaction with EMR functions as voice commands, "e.g. Hey Dragon". However, the actual nature of this was specific to one system at a given time. This interaction will grow and change as EMR design evolves.

Incorporating structured elements along with the free text notes, such as in the EMR data (e.g. MIMIC dataset[43]), would provide another layer of problem complexity. Chiefly, the task transforms from a text-to-text generation task, e.g. dialogue to clinical note summarization, to a multi-modal to multi-modal generation problem, e.g. vitals, past tests, and dialogue to clinical note and orders. The construction of such a dataset is challenging as the nature of mixed free text and structured information per note varies. For example, the same sentence information "Pain level is 3/10" may be represented as entirely free text or in a structured form inserted through a menu. Moreover, there are many correct forms of diagnosis codes and CPT codes per the same information. A diverse dataset would require acquisition of these clinical notes and their underlying mapped structures, requiring exportation by particular EHR. Synthetic creation of such data poses annotation software challenges and would require medical billing expertise.

Comparing the differences of language generation models, including and excluding, EHR inputs and outputs, e.g. ablation experiments with text and structured to text only versus text to text only, can provide insight into problems such as hallucinations and omissions. Furthermore, fine-grained, systematic definition of medical critical and non-critical hallucinations (invented information in the final output) and omissions (e.g. important source transcript content unmentioned in the final output) in evaluation is an area that can benefit from further study.

We hope this dataset will provide a pivotal beginning that can spark the construction and release of more complex future datasets in ambient clinical intelligence. We envision, combined with future work and expansions, this will be one of many progressively complex datasets in the domain that can be used to benchmark progress in the field.

We have provided instructions in the README file in the Figshare repository describing how to process the ACI-BENCH dataset. Examples of processing the data for different summarization evaluations can be found in the code located at the GitHub repository provided below. The data here is intended to be used for benchmarking methods related to clinician-patient dialogue summarization. It should not be used for training models to make medical diagnoses. No patient data was used or disclosed here. Names of the characters were changed (in case content creators used their own information). The gender balance of the entire dataset is roughly equal. Other demographic information was not modeled in this dataset.

Code availability

All code used to run data statistics, baseline models, and evaluation to analyze the ACI-BENCH corpus is freely available at https://github.com/wyim/aci-bench.

Received: 1 May 2023; Accepted: 16 August 2023;
Published online: 06 September 2023

References

1. McDonald, C. J. *et al*. Use of internist's free time by ambulatory care electronic medical record systems. *JAMA internal medicine* **174**, 1860–1863, https://doi.org/10.1001/jamainternmed.2014.4506.
2. Embi, P. J. *et al*. Computerized provider documentation: findings and implications of a multisite study of clinicians and administrators. *Journal of the American Medical Informatics Association: JAMIA* **20**, 718–726, https://doi.org/10.1136/amiajnl-2012-000946.
3. Toll, E. The cost of technology. *JAMA* **307**, 2497–2498, https://doi.org/10.1001/jama.2012.4946.
4. Shultz, C. G. & Holmstrom, H. L. The use of medical scribes in health care settings: A systematic review and future directions. *The Journal of the American Board of Family Medicine* **28**, 371–381, https://doi.org/10.3122/jabfm.2015.03.140224. Publisher: American Board of Family Medicine Section: Original Research.
5. Tran, B. D., Chen, Y., Liu, S. & Zheng, K. How does medical scribes' work inform development of speech-based clinical documentation technologies? a systematic review. *Journal of the American Medical Informatics Association: JAMIA* **27**, 808–817, https://doi.org/10.1093/jamia/ocaa020.
6. Finley, G. *et al*. 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*), 121–128, https://doi.org/10.18653/v1/N18-3015 (Association for Computational Linguistics, New Orleans - Louisiana, 2018).

7. Enarvi, S. *et al*. 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*, 22–30, https://doi.org/10.18653/v1/2020.nlpmc-1.4 (Association for Computational Linguistics, Online, 2020).
8. Krishna, K., Khosla, S., Bigham, J. & Lipton, Z. C. Generating SOAP notes from doctor-patient conversations using modular summarization techniques. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 4958–4972, https://doi.org/10.18653/v1/2021.acl-long.384 (Association for Computational Linguistics, Online, 2021).
9. Zhang, L. *et al*. Leveraging pretrained models for automatic summarization of doctor-patient conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 3693–3712, https://doi.org/10.18653/v1/2021.findings-emnlp.313 (Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021).
10. Michalopoulos, G., Williams, K., Singh, G. & Lin, T. MedicalSum: A guided clinical abstractive summarization model for generating medical reports from patient-doctor conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, 4741–4749 (Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022).
11. Yim, W. & Yetisgen, M. 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*, 10–20, https://doi.org/10.18653/v1/2021.nlpmc-1.2 (Association for Computational Linguistics, Online, 2021).
12. Quiroz, J. C. *et al*. Challenges of developing a digital scribe to reduce clinical documentation burden. *NPJ Digital Medicine* **2**, 114, https://doi.org/10.1038/s41746-019-0190-1.
13. Ben Abacha, A., Yim, W., Fan, Y. & Lin, T. 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*, 2291–2302 (Association for Computational Linguistics, Dubrovnik, Croatia, 2023).
14. Papadopoulos Korfiatis, A., Moramarco, F., Sarac, R. & Savkov, A. PriMock57: A dataset of primary care mock consultations. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 588–598, https://doi.org/10.18653/v1/2022.acl-short.65 (Association for Computational Linguistics, Dublin, Ireland, 2022).
15. Denny, J. C., Miller, R. A., Johnson, K. B. & Spickard, A. Development and evaluation of a clinical note section header terminology. *AMIA Annual Symposium Proceedings* **2008**, 156–160.
16. Podder, V., Lew, V. & Ghassemzadeh, S. SOAP notes. In *StatPearls* (StatPearls Publishing).
17. Yim, W. *et al*. Aci-bench: a corpus for note generation from doctor-patient conversations., *Figshare*, https://doi.org/10.6084/m9.figshare.22494601 (2023).
18. Yim, W., Yetisgen, M., Huang, J. & Grossman, M. Alignment annotation for clinic visit dialogue to clinical note sentence language generation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 413–421 (European Language Resources Association, Marseille, France, 2020).
19. Tiedemann, J. Bitext alignment. In *Synthesis Lectures on Human Language Technologies* (2011).
20. Gliwa, B., Mochol, I., Biesek, M. & Wawer, A. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, https://doi.org/10.18653/v1/d19-5409 (Association for Computational Linguistics, 2019).
21. Jurafsky, D. & Martin, J. H. Speech and language processing: An introduction to speech recognition, computational linguistics and natural language processing. *Upper Saddle River, NJ: Prentice Hall* (2008).
22. Honnibal, M., Montani, I., Van Landeghem, S. & Boyd, A. spacy: Industrial-strength natural language processing in python., *Zenodo*, https://doi.org/10.5281/zenodo.1212303 (2020).
23. Lewis, M. *et al*. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR* **abs/1910.13461** (2019).
24. Lee, J. *et al*. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* **36**, 1234–1240, https://doi.org/10.1093/bioinformatics/btz682 (2019).
25. Gliwa, B., Mochol, I., Biesek, M. & Wawer, A. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, 70–79, https://doi.org/10.18653/v1/D19-5409 (Association for Computational Linguistics, Hong Kong, China, 2019).
26. Beltagy, I., Peters, M. E. & Cohan, A. Longformer: The long-document transformer, https://doi.org/10.48550/ARXIV.2004.05150 (2020).
27. Cohan, A. *et al*. A discourse-aware attention model for abstractive summarization of long documents. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Volume 2 (*Short Papers)* https://doi.org/10.18653/v1/n18-2097 (2018).
28. Lin, C.-Y. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81 (2004).
29. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. *Bertscore: Evaluating text generation with bert* https://doi.org/10.48550/ARXIV.1904.09675 (2019).
30. Sellam, T., Das, D. & Parikh, A. P. *Bleurt: Learning robust metrics for text generation* https://doi.org/10.48550/ARXIV.2004.04696 (2020).
31. Irvin, J. *et al*. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI conference on artificial intelligence* **33**, 590–597 (2019).
32. Soldaini, L. & Goharian, N. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir*, 1–4 (2016).
33. Bubeck, S. *et al*. Sparks of artificial general intelligence: Early experiments with gpt-4 (2023).
34. OpenAI. GPT-4 technical report. https://doi.org/10.48550/ARXIV.2303.08774, Publisher: arXiv Version Number: 3.
35. Hall, J. K. *Teaching and researching: Language and culture* (Routledge).
36. Pitts, M. J. & Gallois, C. Social markers in language and speech. In *Oxford Research Encyclopedia of Psychology*, https://doi.org/10.1093/acrefore/9780190236557.013.300.
37. Clopper, C. G. & Smiljanic, R. Effects of gender and regional dialect on prosodic patterns in american english. *Journal of Phonetics* **39**, 237–245, https://doi.org/10.1016/j.wocn.2011.02.006.
38. Hammami, N. *et al*. Socioeconomic inequalities in adolescent health complaints: A multilevel latent class analysis in 45 countries. *Current Psychology (New Brunswick, N.J.)* 1–12, https://doi.org/10.1007/s12144-022-03038-6.
39. Lampert, T., Kroll, L. E., Kuntz, B. & Hoebel, J. Health inequalities in germany and in international comparison: trends and developments over time. *Journal of Health Monitoring* **3**, 1–24, https://doi.org/10.17886/RKI-GBE-2018-036.
40. Freeborn, L. & Rogers, J. Nonlinguistic factors that affect the degree of foreign accent in second language mandarin. *Studies in Chinese Linguistics* **40**, 75–99, https://doi.org/10.2478/scl-2019-0003.
41. NenkovaAni, PassonneauRebecca & McKeownKathleen. The pyramid method. https://doi.org/10.1145/1233912.1233913. Publisher: ACM PUB27 New York, NY, USA.
42. Mingzhe, L. *et al*. The style-content duality of attractiveness: Learning to write eye-catching headlines via disentanglement. In *AAAI*.
43. Johnson, A. E. W. *et al*. MIMIC-III, a freely accessible critical care database. *Scientific Data* **3**, 160035, https://doi.org/10.1038/sdata.2016.35. Number: 1 Publisher: Nature Publishing Group.

Author contributions

WY developed and created the annotation guidelines, supervised the annotation work, advised on baseline experiments, performed corpus data analysis, and drafted the original manuscript. YF performed baseline experiments, analysis of model performance, and manuscript authorship. AB advised on guideline creation, annotation work, ran baselines, and reviewed and revised the manuscript. NS participated in acquisition of the source data, as well as advised on guideline creation, annotation work, and manuscript review. TL advised on baseline experiments, and reviewed and revised the manuscript. MY advised on the guideline creation, annotation work and baseline experiments, and reviewed and revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-023-02487-3.

Correspondence and requests for materials should be addressed to W.-w.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.