A Typology of Synthetic Datasets for Dialogue Processing in Clinical Contexts

Steven Bedrick[♣] **A. Seza Doğruöz**[♡] **Sergiu Nisioi**[△]

[♣] Division of Informatics & Clinical Epidemiology
Oregon Health & Science University

[♡] LT3, IDLab, Universiteit Gent

[△] HLT Research Center, University of Bucharest

bedricks@ohsu.edu, as.dogruoz@ugent.be, sergiu.nisioi@unibuc.ro

Abstract

Synthetic datasets are used across linguistic domains and NLP tasks, particularly in scenarios where authentic data is limited (or even non-existent). One such domain is that of clinical (healthcare) contexts, where there exist significant and long-standing challenges (e.g., privacy, anonymization, and data governance) which have led to the development of an increasing number of synthetic datasets. One increasingly important category of clinical dataset is that of clinical dialogues which are especially sensitive and difficult to collect. Therefore, they are commonly synthesized. While such synthetic datasets have been shown to be sufficient in some situations, little theory exists to inform how they may be best used and generalized to new applications. In this paper, we provide an overview of how synthetic datasets are created, evaluated and used for dialogue related tasks in the medical domain. Additionally, we propose a novel typology for use in classifying types and degrees of data synthesis, to facilitate comparison and evaluation.

Keywords: health care, synthetic data, clinical conversation datasets

1. Introduction

Synthetic datasets are increasingly used across linguistic domains and NLP tasks, particularly in scenarios where authentic data is limited (or even non-existent). One such domain is that of clinical language, where there are significant and long-standing challenges relating to privacy, anonymization, and data governance, which have led to the development of an increasing number of synthetic datasets (Quintana, 2020; Long et al., 2024).

A particularly fast-growing and relevant domain of clinical language is that of the "clinical dialogue" (i.e., a dialogue between a patient and a health provider, sometimes also referred to as a clinical or medical "conversation"). NLP tools to process such language are being rapidly adopted into practice, most notably in the form of tools for automated generation of clinical documentation (Tierney et al., 2024; Finley et al., 2018) as well as for more traditional dialogue applications (e.g., clinical chatbots). Such dialogues may occur via spoken language, or may be written (taking place via an online discussion forum, a chat-style interface, etc.).

Because of the intense commercial focus in this space, there exists a high level of need for representative dialogue datasets for system development and evaluation (Doğruöz et al., 2023; Adelani et al., 2025). However, due to the extremely sensitive nature of clinical conversations, genuine data is very difficult to obtain, even by the standards of clinical NLP. As such, there are a wide range of synthetic datasets purporting to represent clinical conversations of various types.

While synthetic datasets have been shown to be sufficient in some situations, little theory exists to inform scientific audiences about how these datasets may be best used and generalized to new NLP applications. Towards that end, our goals in this paper are: 1) to provide an overview of how such synthetic datasets are created, evaluated, and used for dialogue-related tasks in the medical domain via a literature review; and 2) to classify these datasets according to a novel typology.

One of the challenges toward this goal is a lack of consensus on what is meant by the word "synthetic" in terms of datasets on dialogues. The canonical definition of synthetic data is that codified by the U.S. Census Bureau in their Statistical Quality Standards (2023), which may be summarized as follows: given a dataset D, generate a new dataset D' in a way that **a)** individual instances in D' are not present in D, **b)** the relevant statistical properties of D' match those of D, and **c)** original instances from D may not be recovered from D'.[1] This definition fits well in the context of a dataset consisting

All authors are corresponding authors [♣♡♠].

[1]This formulation echos that proposed in the earliest literature (Rubin, 1993) on synthetic *microdata*, i.e., the individual data points contained in a dataset: sentences in a corpus, images in a collection, responses from a survey, etc.

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8246–8263
11–16 May 2026.

of structured data. However, as noted by Aggarwal and Yu (2008), it is more challenging to apply this to textual data, for three primary reasons.

First, as compared with traditional structured data, the processes to generate novel linguistic data points are far more complex. It is relatively straightforward to generate synthetic values for scalar or categorical variables by identifying appropriate distributions and then sampling using standard numerical techniques. However, producing language that preserves the *semantic* and *discursive* properties of the original necessitates more sophisticated modeling and sampling techniques. Secondly, in the context of a structured dataset, the phrase "relevant statistical properties" can be operationalized in terms of univariate or joint probabilities of the constituent variables. However, for linguistic data, the possibility space is far larger, and relying on simple statistical methods (e.g. perplexity) is unlikely to have satisfactory results. That is to say, the final dataset is unlikely to be truly "equivalent" to the original in linguistic terms (see §3). Finally, the question of re-identifiability (constraint "c" from the above definition) is subtle and complex to assess in the context of linguistic data.

Another challenging aspect of synthetic data is that artificiality is not a binary property. All datasets, being curated and abstracted representations of some sampled reality, are inherently synthetic to some degree. This process necessarily involves intervention around questions of inclusion/exclusion, representation, and structure. As Bowker (2005) put it, "raw data is both an oxymoron and a bad idea; to the contrary, data should be cooked with care." However, while all datasets may be synthetic, some datasets are more synthetic than others: a dataset may be "more" or "less" synthetic, depending on its construction and the point of view of evaluation. Literature on synthetic data distinguishes between datasets that are "fully synthetic," being those that contain no original, genuine microdata, and "partially synthetic," in which only certain variables are perturbed (Reiter, 2003). In terms of *structured* medical data, Gonzales et al. (2023) and Murtaza et al. (2023) provide comprehensive discussions on this and other relevant taxonomies.

How can this framework, originally designed for numerical data, be applied to linguistic data, which focuses on meaning and grammar rather than statistics?

Exploring this question is the overall goal of our paper. In line with this goal, first, we review various ways in which synthetic language datasets are produced (§2). Next, we more formally define our terms to explore this continuum of synthesis (§3), and propose a typology that captures these subtleties in a practical and actionably useful way (§3.1). We also report the results of a literature review of datasets relating to clinical dialogues (§4.1) and demonstrate how our proposed typology describes existing datasets. Finally, we close with a discussion that extends beyond the microdata-centric view of our paper, to explore how realism and synthesis may relate to contextual aspects of a dataset (§4.2).

2. Means of Production

Synthetic data production methods can be categorized into "process-driven"[2] and "data-driven" methods. Under the former, whatever generative process is carrying out the synthesis is informed by human knowledge and/or by information that is exogenous to the dataset itself. The latter describes methods that use properties of the dataset itself to derive the generative model (Goncalves et al., 2020; Murtaza et al., 2023).

This framing emphasizes a description of the mechanical process by which the synthesis is done. While important, it does not describe *what* is being synthesized, nor does it address the degree to which a given dataset should be considered synthetic. This omission is problematic when considering synthetic linguistic data, as there are numerous *aspects* (e.g., topic, genre/register, grammatical or discursive structure, specific aspects of semantic content, individual lexical choices, etc.) of the data that might (or might not) be synthesized. A given dataset might involve synthesis along one of these domains, while leaving other domains un-perturbed. To illustrate this issue, consider the question of how synthetic clinical conversational datasets are created.

From our review, we find that such datasets are created using one of several methods. First, they can be created by taking "genuine" data (i.e., records of authentic patient-provider conversations that actually occurred in the "real world"), and perturbing it in some way. This perturbation may be intended to obscure identifiable information about actual patients (Negash et al., 2023), to produce a larger volume of comparable data than originally existed (e.g. Modersohn et al. (2022)), or as a form of data augmentation (Abdollahi et al., 2021).

Second, a conversational dataset can be created *de novo* via a human-driven manual process. In medical education, it is very common for human actors to collaborate with clinical trainers or trainees to simulate various kinds of clinical interactions (Rutherford-Hemming et al., 2024). Recordings of such simulated encounters, as well as other artifacts such as encounter notes, etc., can serve as rich sources of data that are free from concerns about patient privacy. While such secondary use is

[2]Sometimes referred to as "theory-" or "knowledge-driven."

rarely the *goal* of simulations when done in didactic contexts, they can also be performed specifically for the purpose of generating a re-usable dataset (Fareez et al., 2022; Yim et al., 2023).

Alternatively, human annotators (trained or otherwise) may be tasked with *writing* an entire end-to-end conversation given pre-existing parameters and guidelines (Shi et al., 2023; Ben Abacha et al., 2023). In such scenarios, the resulting dialogue is not the natural result of humans interacting. It is essentially "screenwritten," with the quality and utility of the resulting artifact depending heavily on the knowledge, experience, and skill of the author.

In simulated scenarios involving humans, the pragmatic and paralinguistic components of the resulting dialogues may be less representative of what would be seen in a true naturalistic exchange. Similarly, we would expect a "screenwritten" scenario imagined by a single author to be less representative than a simulation involving multiple humans interacting in a more naturalistic way. However, depending on the intended use of the resulting data, this may not be an issue. For example, if the primary content of interest is the semantic or lexical contents of the exchange, as opposed to the pragmatic or discursive contents, such a dataset may be entirely adequate. If, however, the analytical question of interest depends on discourse features, a synthetic dataset of this sort would likely be insufficiently representative.[3]

Finally, simulated datasets may be created *de novo* via a mechanical process, e.g. using a natural language generation (NLG) model (Frei and Kramer, 2023). This may take the form of prompted generation (i.e., an LLM being prompted to write a note or simulate a dialogue given certain criteria) or in some cases by literally having two LLMs "talk" to one another in a simulated dialogue, in an effort to increase the variety of the resulting output (Gray et al., 2024).

Human-simulated scenarios have a long history of use in medical education and evaluation (Issenberg et al., 1999; McGaghie et al., 2010), resulting in a large and well-validated family of methodologies designed to ensure that the resulting interactions are sufficiently representative for their intended purpose (typically clinical training and assessment). Despite early and longstanding interest (Starkweather et al., 1967), computer-driven *de novo* dialogue generation has been relatively rarely used in clinical education. However, modern systems are beginning to be evaluated for use in medical training (Awada et al., 2024; Holderried et al., 2024), leading to the inevitable creation of an additional form of synthetic dataset, in which a human role-plays either the patient or the provider against the LLM playing the opposite role (Campillos-Llanos et al., 2021; Abercrombie and Rieser, 2022; Zhang et al., 2023)

Given this variety of means of generation, we argue that it is inaccurate to think of a dataset as being either "synthetic" or "real". Instead, we argue that any given dataset exists as a point along a continuum between the two poles. After all, even the most "real-world" datasets are inevitably the result of numerous manual interventions and decisions (Gitelman, 2013; Passi and Sengers, 2020; Sambasivan et al., 2021).

Similarly, we argue that an otherwise-genuine dataset that has been "anonymized" is, in some sense synthetic, with the degree to which this is the case depending on the method of anonymization in question. At one end of the spectrum, imagine a scheme in which identifiers are replaced with a class token of some kind (e.g. `PATIENT_NAME`, `DAY_OF_WEEK`, etc.) in a purely deterministic and uniform manner. At the other end of that spectrum, consider a method in which identifiers are replaced with statistically-plausible alternative values ("pseudonymization")[4]

In both cases, the process of applying such a method results in the creation of a new and distinct dataset. We argue, however, that the second case is "more synthetic" in that there are two levels of synthesis in play: the simple alteration of the original un-modified data, in combination with some *designed process* informing the pseudonymization. This process must itself necessarily bring with it working assumptions, and may also depend on yet additional sources of data, further distancing the resulting synthetic dataset from whatever grounding in reality it may have had earlier in its development.

3. What Do We Mean By "Synthetic"?

In response to the complexities surrounding the question of whether a given dataset is "synthetic" or not, we propose a conceptual framework whose aim is to provide structure and clarity when describing datasets that have a synthetic component, and to facilitate meaningful comparisons. For our purposes, we define a *dataset* as a collection of information-bearing objects that have been purposefully assembled or otherwise grouped, with the goal of serving a particular representational function, with the intention of being used in analysis (broadly defined) of some kind. See Appendix A for a detailed discussion of this definition and its constituent elements.

How does our above definition fit with *synthetic* data? A dataset is, by definition, a human-generated artifact. At all stages of collection, cura-

[3] See section 4.2 for additional discussion.

[4] See Uzuner et al. (2007) for a detailed description of such a method, as well as an overview on deidentification in non-dialogue clinical text in general.

	Description	Human	Machine
Type 1	No intervention	-	-
Type 2	Altering existing microdata	Altering an existing "real" dialogue, according to a specification	Masking names, translating into Japanese, etc.
Type 3	Generating entirely new microdata	Writing a new dialogue de novo, according to a narrative prompt	LLM-generated dialogue turns

Table 1: Comparison of human and machine microdata interventions

tion, and processing, choices are made that separate the resulting dataset from its originating context. As such, we further define a "naturalistic" dataset as one that fulfills the terms of our above definition, along with the additional constraint that the microdata contained in the dataset are **a)** intended by the designers to be *faithful representations of some genuine originating phenomenon*, and **b)** are understood to be such by the users of the dataset.

For example, consider the MedDialog corpus of patient-provider dialogues taken from an online medical consultation website (Zeng et al., 2020). The designers of this dataset understood and believed these dialogues to represent exchanges that actually took place between human patients and human providers who actually existed. While they may have applied criteria to which dialogues were included (perhaps focusing on a particular category of patient, or clinical specialty), the lexical, semantic, pragmatic properties of the language and sociolinguistic aspects of the clinical dialogue were left un-modified in the hope of comprising a maximally representative sample. In turn, downstream *users* of this dataset would understand its contents to be representative of exchanges that actually took place (i.e., to be "real messages").

In comparison, consider the `ACI` subset of the ACI-bench dataset (Yim et al., 2023), which consists of transcribed exchanges between an actual doctor and a volunteer "patient," acting out a scenario based on symptom-driven prompts. The exchanges in such a dataset are "real" in the sense that they did, in fact, take place, and occurred between actual humans. However, they are not "real" in that they arose from an artificial scenario, as opposed to the actual real-world domain that they are attempting to represent.

The creators of ACI-bench, of course, attempted to design their scenarios to be as "realistic" as possible, and gave careful consideration to their design. Their goal was to create a dataset whose synthetic nature did not affect those aspects of its form or content to an extent that compromised the dataset's fundamental *validity*[5] for the analyses of the sort that the dataset was intended by its authors and users to support. Decisions about whether, when, and how to *use* such a dataset typically hinge on the extent to which this assumption holds for one's particular purpose.

Finally, consider the DoPaCo dataset (Chen et al., 2023a). Here, the authors began with transcribed and de-identified genuine doctor-patient conversations, to which they had extremely restricted and time-limited access, and which could not be re-distributed. They used this corpus to fine-tune a pre-trained dialogue-generation Transformer model, and then used this model to generate a large corpus of synthetic dialogues, which they used for the rest of their work. One may argue that such a dataset is in some sense *meaningfully or usefully representative* of a "real" dataset of doctor-patient conversations, according to any number of rubrics.[6] But, it is also equally true that such a dataset, however *statistically* similar it may be to its original training data, is not comprised of "real" dialogue data in the same sense as ACI-bench or MedDialog. Rather than representing true discursive choices made by humans interacting with one another, the dialogues in the DoPaCo dataset reflect that model's statistical approximation of the discursive choices represented in its training data.

This question of "real-ness" is an entirely separate question from whether such a dataset is *useful* or *sufficient* for a given purpose. We are again using "real" as a *descriptive* rather than a *normative* term, and of course even highly synthetic datasets have many uses. As such, we are emphatically *not* arguing that datasets that are more naturalistic are

[5]Here we use "validity" in the specific senses of construct and measurement validity; see Brewer and Crano (2014) for an overview of these terms, and Jacobs and Wallach (2021) for a discussion of how these concepts apply in machine learning research.

[6]Given that it was produced via a sequence-to-sequence model, one may accurately and literally describe it as consisting of a set of samples drawn from a statistical model of the originating "real" dataset; as such, there exist statistical measures that we may use to quantify the degree to which such samples are or are not representative (in a purely statistical sense) of the originating data, such as perplexity and KL-Divergence. However, we also note that there is a long-standing and robust literature (Galliers and Spärck Jones, 1993; Iyer et al., 1997; Ito et al., 1999) on the *limitations* of such purely statistical measures in terms of their external and ecological validity (Brewer and Crano, 2014; Egami and Hartman, 2023).

inherently and categorically superior in some way to datasets that are more synthetic, though there may indeed exist scenarios where this is the case; that question is orthogonal to our purpose in the present work. Finally, we note that the synthetic datasets we describe in this paper are uniformly accompanied by more or less detailed evaluations by their creators, seeking to investigate the degree to which they are or are not sufficient for specific applications.

3.1. A Typology for Human/Machine Microdata Interventions

We propose a three-level typology of "types" of intentional intervention to the *content* of a dataset's microdata (illustrated in Table 1). Our typology's categories are inspired by the perturbation-generation dichotomy described by Domingo-Ferrer (2008), adapted and expanded in terms appropriate to the complexities of linguistic datasets (as opposed to traditional statistical datasets). Additionally, we observe that the interventions involved in producing a synthetic linguistic dataset may be carried out by either humans or machines (or both) during the course of a dataset's creation. In practice, many synthetic datasets involve a mix of both, and adequately capturing the ways in which a given dataset is "synthetic" requires that both aspects be described. As such, our typology involves categorizing a dataset along dimensions of both the *human* and *machine* involvement (Table 1). This increased expressiveness further helps break down the false binary of "synthetic" vs. "real," and allows clearer descriptions of a given dataset.

- Type 1. No intervention (i.e., there is no automatic or manual process of alteration, and data reflect naturally-occurring linguistic exchanges).

- Type 2. Existing microdata is perturbed according to a clear specification (e.g., anonymization, paraphrasing, redaction). These perturbations are traceable, with the synthetic data maintaining a direct lineage to the source data. Human- or machine-mediated feedback may guide these alterations, but the structure of the source data is preserved to a recognizable degree.

- Type 3. New microdata are produced *de novo* via a generative process of some kind, to substitute for "real" microdata. The generation may be conditioned on real data, and may also involve exogenously-provided knowledge in the form of prompts or templates. Iterative refinement involving human feedback or reinforcement learning (e.g., via RLHF[7] or human simulations) falls under this category when it results in the creation of fundamentally new data samples.

In our typology, these levels are not mutually exclusive. It is possible for a dataset to include both Type 2 and Type 3 interventions. Similarly, a dataset may involve multiple interventions of the same or a different type, focused on different aspects of its contents.

Consider a dataset of "screen-written" patient-provider interactions, in which the dialogues' contents and discursive structure are entirely imagined by humans in response to a prose case description (for example, the MTS-Dialog dataset (Ben Abacha et al., 2023)). As the representational goal of this dataset is to contain "realistic" clinical dialogue, to be used in place of otherwise unavailable genuine dialogues, we consider this to be a "Human Type 3, Machine Type 1" dataset.

However, if the *goal* of the dataset had been to consist of simulated dialogues (as opposed to the actual goal, to wit, using *simulated* dialogues as *representative stand-ins* for *genuine* dialogues), we would instead consider the human-written dialogues to be non-synthetic primary microdata, and categorize the human component of this dataset under "Type 1."

This example highlights the value of our typology. It facilitates clarity about both the representational goals and the actual content of a given dataset. By doing so, we make it easier for future producers of datasets to communicate explicitly about these aspects of their creations, and for consumers of datasets to reason concretely about whether their intended use aligns with the dataset's capabilities and contents.

3.2. In Search of Synthetic Dialogues

Our typology's design was informed by a literature review whose goal was to survey the landscape of synthetic biomedical dialogue datasets. We defined our clinical domain of interest as comprising text capturing either spoken or written language generated by patients and/or health providers, intended for use as part of clinical care, patient information-seeking, etc. Our goal was to identify published datasets of clinical dialogues, in spoken or written genre, and either synthetic or natural in nature.

Included publications could be specifically focused on the production and evaluation of their datasets, or could produce datasets as secondary research artifacts. In the latter case, we required

[7] Reinforcement Learning from Human Feedback (Christiano et al., 2017).

that there be an intention on the part of the authors that their datasets be re-used by others (i.e., their data were not simply made re-usable for purposes of replicability). We used keyword-based and index-term search strategies, supplemented by large language models (LLMs) to assist in filtering and classifying retrieved papers. See Appendix B for a detailed description of our approach and Appendix D for the verbatim LLM prompts. Table 2 summarizes the results of our structured search strategies across PubMed, the ACL Anthology, and dblp.

From a total of 1,626 papers (679 from PubMed, 591 from ACL Anthology, and 356 from dblp), successive filtering steps (including LLM-assisted classification and manual review) reduced the pool to 25 included papers (4 from PubMed, 11 from ACL Anthology, and 10 from dblp). After removing overlapping entries, we arrived at 20 unique papers.

From a scientific standpoint, we note that clinical dialogue-oriented datasets are more variable in terms of their linguistic domains, intended uses, structure, means of collection, methods of synthesis, etc. than the larger but comparatively more homogeneous body of literature describing synthetic electronic health record (EHR) notes. As such, they provide an informative corpus of examples to use in attempting to understand and categorize different approaches to synthesis. In practice, however, our typology (and this paper's broader discussion of synthesis) would be applicable to datasets containing synthetic notes or other forms of clinical language.

Given our focus on dialogues, we excluded the following categories of dataset from our analyses in this paper:

1. Published datasets whose primary content was medical in nature but did *not* consist of dialogues, e.g., electronic health record (EHR) notes, structured EHR data, etc. This resulted in the exclusion of most existing work on data augmentation and synthetic medical data more generally, and on clinical language in particular. By far the most frequently-seen category of synthetic clinical language dataset is that of EHR notes.

2. Datasets where the synthetic component was in the form of structured annotations (e.g., dialogue act classification, concept/entity labels, etc.), narrative summaries, or other secondary metadata, as opposed to the primary dialogue itself.

3. Datasets that *use* genuine dialogues, but whose artifacts of interest and analysis are synthetically-produced secondary products (such as summaries or clinical notes) generated automatically from those original genuine dialogues.

We also excluded datasets that are not themselves presenting or simulating dialogues as such, but which *are* intended to produce artifacts for use in training systems whose modality of use is conversational in nature. For example, Kang et al. (2024) describe a process for transforming curated biomedical data into pseudo-conversational question-answer pairs for use in instruction tuning a downstream LLM. While the intended *purpose* of such a system would be to support conversational interaction between a human and an LLM, *the data itself* are not conversational in nature nor does it attempt to characterize the key linguistic features of a dialogue-based interaction.

As such, papers describing such datasets were of secondary interest, and we have excluded them from our review. We note that, alongside synthetic EHR notes, synthetic QA pairs are also a very prevalent form of synthetic clinical dataset.

Another category that we ultimately excluded from our review are papers that generated synthetic dialogue data as a step towards some *other* primary analysis, but which did not release their resulting datasets as artifacts for use by others. One prototypical example of this category of work includes that of Yosef et al. (2024), who carried out a series of experiments designed to assess the ability of LLMs to simulate patient/therapist interactions. While they do produce synthetic dyadic interactions as part of their work, their primary interest is using them as inputs to another LLM-based analysis designed to determine the ability of LLMs to differentiate between styles and forms of therapy. A second example of this category of work would be that of Gray et al. (2024), who used LLMs to simulate patient-provider counseling discussions in a neonatology context and conducted a very thorough analysis in order to assess their content and utility from a clinical perspective.

In both cases, the authors did not publish their resulting simulated interactions, and in both cases, producing these simulated exchanges was not the primary (or even secondary) point of their study. Notably, if they *had* published their linguistic datasets, the resulting corpora from both papers would be usefully described by our typology (both would be Human-1/Machine-3, in that all linguistic microdata were generated entirely "from scratch" by the machine with no human involvement).

4. Discussion

4.1. Synthetic Data and Varieties of Task

As discovered in our review, synthetic datasets in the clinical domain are used for a variety of NLP related tasks such as: Information Extraction (Zhang et al., 2020), summarization of the clinical dialogues (Joshi et al., 2020; Chintagunta et al., 2021), generating notes from conversations (Yim et al., 2020), dialogue with custom personality traits for training

Human \ Machine	Type 1 **No intervention**	Type 2 **Alteration**	Type 3 **Generation**
Type 1 **No Intervention**	Zeng et al., (2020)	Li et al., (2021) Liu et al., (2022) Xu et al., (2024) Binici et al., (2025)	Chen et al., (2023a) Zhang et al., (2023) Han et al., (2024) Moser et al., (2024) Nishiyama et al., (2024) Wang et al., (2024)
Type 2 **Alteration**	Chen et al., (2023a)	Shi et al., (2023)	
Type 3 **Generation**	Liu et al., (2021) Papadopoulos Korfiatis et al., (2022) Fareez et al., (2022) Ben Abacha et al., (2023) Yim et al., (2023)		Zheng et al., (2023) ALMutairi et al., (2024)

Figure 1: A selection of clinical dialogue datasets classified according to different degrees of human-machine interventions.

purposes (Han et al., 2024), training data for ASR systems (Fareez et al., 2022), and named entity recognition (Frei and Kramer, 2023).

Language Diversity in Datasets: As is commonly observed across NLP tasks, English was the most prominent language among the datasets we identified (e.g. Joshi et al. (2020); Chintagunta et al. (2021); Abercrombie and Rieser (2022); Wang et al. (2023a); Ben Abacha et al. (2023); Yim et al. (2023)). In addition, there is a growing amount of research for Chinese (Mandarin) datasets (Zhang et al., 2020; Shi et al., 2023) as well as in Arabic (ALMutairi et al., 2024), Korean (Han et al., 2024), and German (Röhrig et al., 2022; Frei and Kramer, 2023).

Type of Medical Domains: Most datasets we were able to retrieve are not targeted for a specific type of disease in the clinical domain, and instead attempted to more broadly represent written clinical notes or patient-provider dialogues from generic clinical settings (e.g. an outpatient visit). However, we did encounter a handful of synthetic datasets focusing on specific clinical specialties, including cardiology (Zhang et al., 2020), gastroenterology (Liu et al., 2022), psychology (Liu et al., 2021; Zheng et al., 2023; Han et al., 2024), and obstetrics & gynecology (Wang et al., 2023b).

4.2. On Beyond Microdata

To this point, we have limited our discussion to the lowest and most granular level of a dataset's contents, its microdata. Beyond what we have presently described, there is also another dimension of synthesis that we believe to be relevant. A linguistic dataset includes information at multiple levels of representation, including lexical (which words are used), syntactic (how are sentences constructed), and semantic (what meaning is being expressed). In the case of a dialogue-oriented dataset, there may also be pragmatic and discursive information (who is speaking, in what register(s) and with what turn-taking structure, etc.). Additionally, there may be extra-linguistic information represented in such a dataset, for example, the larger social and phenomenological context from which the data are meant to have arisen. Interventions or other synthetic data generation processes may be targeted at, or otherwise affect, all, some, or none of these layers.

There are also other factors that play important roles in the "realism" of a dataset. Recall the MTS-Dialog dataset of Ben Abacha et al. (2023), which consists of "dialogues" that were imagined and "screen-written" *de novo* by humans (in this case, trained clinicians) based on short clinical case vignettes. Now, imagine a hypothetical second dataset ("dataset *B*") in which, instead of being screen-written dialogue, the discourse samples were in fact the result of transcribed conversations taking place between two actors, posing as a doctor-patient dyad and improvisationally acting out a scenario based on the same case vignettes; the dataset described by Fareez et al. (2022) follows this method of creation, as does the `ACI` subset of the ACI-bench dataset (Yim et al., 2023).

In both of these examples, we would consider the resulting dataset to be synthetic, in the sense that neither one reflects the language produced by actual "real world" participants in the true interaction that the dataset is purporting to represent (i.e., a clinical encounter). Our typology in its present form would categorize both as Type 3 datasets along their human dimensions,[8] and in both cases we can safely say that the context that gave rise to the linguistic information contained in the dataset is surely more artificial than in a hypothetical *third* dataset ("dataset *C*") containing "genuine" conversations between patients and providers captured "in the wild" (a Type 1 dataset, according to our typology).

However, this thought experiment illustrates a limitation of our present typology. The linguistic artifacts in "dataset B" (e.g., the utterances themselves, along with their associated paralinguistic and discursive content) were produced by actual humans in conversation, and could certainly be thought of as being "less synthetic" than the utterances in MTS-Dialog, which were made up from whole cloth. In both scenarios, the semantic content of the utterances may be quite similar, as might the degree to which they accurately reflect the vignettes that formed their seeds. The variation would involve

[8] As the microdata are being "made up" (as opposed to having naturally arisen or being modified *in situ*).

the above-mentioned pragmatic and discursive aspects of the data, as a side effect of the difference in the method of generation. Our typology currently lacks a notion of, for lack of a better word, *modality*: in this case, "imagined" vs. "transcribed" spoken language. It also does not distinguish between degrees of synthesis along these dimensions.

Another area that our typology does not capture is the relationship between synthesis and a dataset's sociolinguistic context (Nguyen et al., 2016). Consider a hypothetical dataset of transcribed English-language patient-provider conversations taking place during genuine outpatient visits in the United States (i.e., a Type 1, "real" dataset in our typology). Imagine, now, that the dataset were run through a machine translation (MT) algorithm, with the intent of producing a French-language dataset for use in evaluating algorithms designed to process Francophone clinical encounters. According to our typology, this is now a "Type 2" dataset: the dataset's microdata are still clearly derived from "real" microdata, but have been perturbed in some systematic way.

Depending on the quality of the machine translation system, the resulting dataset may be "equivalent" in terms of its lexical and syntactic representativeness (due to the similarity between English and French), and may also be equivalently useful in terms of its semantic content (i.e., the set of clinical conditions and events being described). In other words, the impact of the synthesis may be minimal along those dimensions. However, the *pragmatic* and *discursive* content may be inappropriate, as norms of clinical communication and patient expectations vary significantly between cultures and countries.[9] As such, even assuming a very high-quality translation, the dataset is likely no longer as valid, in terms of its linguistic representativeness, as it may have been before being made synthetic. This is likely the case to a degree that is much larger than might be expected by a linguistically naïve researcher equipped with an MT algorithm.[10]

Additionally, consider that the extra-linguistic factors represented in our hypothetical dataset may be severely impacted by this sort of perturbation. In this particular example, the translated conversation may indeed be adequately representative in terms of its linguistic and clinical content, but its contextual representativeness may be impaired. A patient-provider discussion about the management of end-stage kidney disease in American English will include discussion of lab values, diet, etc., all of which may translate into a French context with adequate fidelity. However, the parts of the conversation involving the logistics of insurance coverage for dialysis, or of organizing a social worker to assist in transportation to and from dialysis sessions, would differ wildly between the two contexts.

This scenario is inspired by an actual example we encountered during our literature review, in which ACI-bench (a Human Synthetic Type 3 dataset) was used in a one-shot setting to prompt LLMs to generate equivalent dialogues in the Najdi Arabic dialect, with the goal of creating a dataset for use in cultural adaptation of an LLM (ALMutairi et al., 2024). According to our typology, the resulting data is machine-generated while also being originated by a human-driven *de novo* synthesis process and falls under Human Type 3, Machine Type 3. As LLMs are increasingly deployed at scale, hybrid synthetic datasets of this type will likely become more common. However, the validity and representativeness of such datasets must be carefully assessed.

5. Conclusion

Although synthetic datasets are widely used for dialogue generation, the boundaries of what makes a dataset synthetic are not always clear, and they vary among fields. Focusing on the healthcare domain and only on dialogue data, we propose a typology to clarify the types of synthetic datasets in terms of the human and machine interventions.

Academic literature on synthetic data focuses heavily on the mechanics of *how* synthetic data elements are generated, and our typology focuses instead on the role that the synthesis plays in the overall "shape" of the dataset. Our typology provides authors with the ability to speak in a more clear and granular way about what aspects of their dataset are synthetic, to what degree, and from what source (human vs. machine).

Our literature review revealed a clear pattern of increased prevalence of fully-synthetic (Type 3) datasets, particularly of machine-generated dialogues following the widespread availability of LLMs in 2023. Given this, our work's broader perspective on syntheticity will help researchers better assess and contextualize the validity and representativeness of these datasets.

The relationship between synthesis and a dataset's larger contextual factors remains currently unexplored. As future work, we hope to expand our work to capture realism and synthesis along

[9] We note that, even within societies, norms of clinical communication vary greatly across numerous social and demographic factors, as well as between clinical specialties. This is one of the reasons why we are particularly wary of synthetic data in the context of clinical dialogues.

[10] Despite technical advances, translation remains, of course, notoriously challenging (Anonymous, 1968) and finicky (Carlson et al., 2007). Clinical contexts bring numerous specialized challenges for the use of MT (Mehandru et al., 2022).

non-linguistic and contextual factors.[11]

6. Limitations

Our typology provides a structured framework for describing synthetic dialogue datasets and we summarize here some of its limitations discussed broadly in §4.2.

First, the categories of microdata interventions do not capture variations in modality (i.e., whether dialogues are imagined by a single author, simulated through human interaction, or transcribed from naturalistic speech). As such, pragmatic and discursive elements may differ significantly between these modalities.

Second, our work does not account for the sociolinguistic context of the synthetic data. A dataset may be linguistically equivalent after MT, yet fail to reflect cultural norms of clinical communication, potentially reducing its validity.

Currently, the ability to characterize amounts of synthesis along these dimensions is limited, and at the present time, we consider these questions to be left for future work.

Acknowledgments

Sergiu Nisioi is supported by the project "Romanian Hub for Artificial Intelligence - HRIA", Smart Growth, Digitization and Financial Instruments Program, 2021-2027, MySMIS no. 351416. Steven Bedrick is supported by the National Library of Medicine of the National Institutes of Health under award number 5R01LM011934-10 ("Semi-structured Information Retrieval in Clinical Text for Cohort Identification").

7. Bibliographical References

Mahdi Abdollahi, Xiaoying Gao, Yi Mei, Shameek Ghosh, Jinyan Li, and Michael Narag. 2021. Substituting clinical features using synthetic medical phrases: Medical text data augmentation techniques. *Artificial Intelligence in Medicine*, 120:102167.

Gavin Abercrombie and Verena Rieser. 2022. Risk-graded safety for handling medical queries in conversational AI. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 234–243, Online only. Association for Computational Linguistics.

David Ifeoluwa Adelani, A. Seza Doğruöz, Iyanuoluwa Shode, and Anuoluwapo Aremu. 2025. Does generative AI speak Nigerian-Pidgin?: Issues about representativeness and bias for multilingualism in LLMs. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1571–1583, Albuquerque, New Mexico. Association for Computational Linguistics.

Charu C. Aggarwal and Philip S. Yu. 2008. A General Survey of Privacy-Preserving Data Mining Models and Algorithms. In Ahmed K. Elmagarmid, Amit P. Sheth, Charu C. Aggarwal, and Philip S. Yu, editors, *Privacy-Preserving Data Mining*, volume 34, pages 11–52. Springer US, Boston, MA.

Mariam ALMutairi, Lulwah AlKulaib, Melike Aktas, Sara Alsalamah, and Chang-Tien Lu. 2024. Synthetic Arabic medical dialogues using advanced multi-agent LLM techniques. In *Proceedings of The Second Arabic Natural Language Processing Conference*, pages 11–26, Bangkok, Thailand. Association for Computational Linguistics.

Anonymous. 1968. Mokusatsu: One word, two lessons. *NSA Technical Journal*, 13(4):95–100.

I. A. Awada, A. M. Florea, and A. Scafa-Udriște. 2024. An e-learning platform for clinical reasoning in cardiovascular diseases: a study reporting on learner and tutor satisfaction. *BMC Medical Education*, 24(1):984.

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, Dubrovnik, Croatia. Association for Computational Linguistics.

Kuluhan Binici, Abhinav Ramesh Kashyap, Viktor Schlegel, Andy T. Liu, Vijay Prakash Dwivedi, Thanh-Tung Nguyen, Xiaoxue Gao, Nancy F. Chen, and Stefan Winkler. 2025. MEDSAGE: enhancing robustness of medical dialogue summarization to ASR errors with llm-generated synthetic dialogues. In *AAAI*, pages 23496–23504. AAAI Press.

Geoffrey C. Bowker. 2005. *Memory Practices in the Sciences*. Inside Technology. MIT Press, Cambridge, Mass.

[11] The extent to which there exists a valid dichotomy between "linguistic" and "non-linguistic" content in a dataset is surely up for debate.

Marilynn B Brewer and William D Crano. 2014. Research design and issues of validity. In Harry T Reis and Charles M Editors Judd, editors, *Handbook of Research Methods in Social and Personality Psychology*, pages 11–26. Cambridge University Press, New York.

Suzanne Briet. 1951. *Qu'est-ce que la documentation?* Collection de documentologie, 1. Éditions documentaires, industrielles et techniques, Paris.

Michael K. Buckland. 1991. Information as thing. *Journal of the Association for Information Science and Technology*, 42(5):351–360.

Leonardo Campillos-Llanos, Catherine Thomas, Éric Bilinski, Antoine Neuraz, Sophie Rosset, and Pierre Zweigenbaum. 2021. Lessons learned from the usability evaluation of a simulated patient dialogue system. *Journal of Medical Systems*, 45(7):69.

Robert V Carlson, Nadja H van Ginneken, Luisa M Pettigrew, Alan Davies, Kenneth M Boyd, and David J Webb. 2007. The three official language versions of the Declaration of Helsinki: What's lost in translation? *Journal of medical ethics*, 33(9):545–548.

Siyuan Chen, Colin A. Grambow, Mojtaba Kadkhodaie Elyaderani, Alireza Sadeghi, Federico Fancellu, and Thomas Schaaf. 2023a. Investigating the Utility of Synthetic Data for Doctor-Patient Conversation Summarization. In *Interspeech 2023*, pages 2338–2342.

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

Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2021. Medically aware GPT-3 as a data generator for medical dialogue summarization. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 66–76, Online. Association for Computational Linguistics.

Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

A. Seza Doğruöz, Sunayana Sitaram, and Zheng Xin Yong. 2023. Representativeness as a forgotten lesson for multilingual and code-switched data collection and preparation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5751–5767, Singapore. Association for Computational Linguistics.

Josep Domingo-Ferrer. 2008. A Survey of Inference Control Methods for Privacy-Preserving Data Mining. In Ahmed K. Elmagarmid, Amit P. Sheth, Charu C. Aggarwal, and Philip S. Yu, editors, *Privacy-Preserving Data Mining*, volume 34, pages 53–80. Springer US, Boston, MA.

Naoki Egami and Erin Hartman. 2023. Elements of External Validity: Framework, Design, and Analysis. *American Political Science Review*, 117(3):1070–1088.

Faiha Fareez, Tishya Parikh, Christopher Wavell, Saba Shahab, Meghan Chevalier, Scott Good, Isabella De Blasi, Rafik Rhouma, Christopher McMahon, Jean-Paul Lam, et al. 2022. A dataset of simulated patient-physician medical interviews with a focus on respiratory cases. *Scientific Data*, 9(1):313.

Gregory Finley, Erik Edwards, Amanda Robinson, Michael Brenndoerfer, Najmeh Sadoughi, James Fone, Nico Axtmann, Mark Miller, and David Suendermann-Oeft. 2018. An automated medical scribe for documenting clinical encounters. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 11–15, New Orleans, Louisiana. Association for Computational Linguistics.

Johann Frei and Frank Kramer. 2023. Annotated dataset creation through large language models for non-english medical nlp. *Journal of Biomedical Informatics*, 145:104478.

J.R. Galliers and K. Spärck Jones. 1993. Evaluating natural language processing systems. Technical Report UCAM-CL-TR-291, Computer Laboratory, University of Cambridge.

Lisa Gitelman. 2013. *"Raw Data" Is an Oxymoron*. The MIT Press.

Andre Goncalves, Priyadip Ray, Braden Soper, Jennifer Stevens, Linda Coyle, and Ana Paula Sales. 2020. Generation and evaluation of synthetic patient data. *BMC Medical Research Methodology*, 20(1):108.

Aldren Gonzales, Guruprabha Guruswamy, and Scott R. Smith. 2023. Synthetic data in health care: A narrative review. *PLOS Digital Health*, 2(1):e0000082.

Megan Gray, Austin Baird, Taylor Sawyer, Jasmine James, Thea DeBroux, Michelle Bartlett, Jeanne Krick, and Rachel Umoren. 2024. Increasing Realism and Variety of Virtual Patient Dialogues for Prenatal Counseling Education Through a Novel Application of ChatGPT: Exploratory Observational Study. *JMIR medical education*, 10:e50705.

Ji-Eun Han, Jun-Seok Koh, Hyeon-Tae Seo, Du-Seong Chang, and Kyung-Ah Sohn. 2024. PSY-DIAL: Personality-based synthetic dialogue generation using large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13321–13331, Torino, Italy. ELRA and ICCL.

Friederike Holderried, Christian Stegemann-Philipps, Anne Herrmann-Werner, Teresa Festl-Wietek, Martin Holderried, Carsten Eickhoff, and Moritz Mahling. 2024. A Language Model-Powered Simulated Patient With Automated Feedback for History Taking: Prospective Study. *JMIR medical education*, 10:e59213.

S. B. Issenberg, W. C. McGaghie, I. R. Hart, J. W. Mayer, J. M. Felner, E. R. Petrusa, R. A. Waugh, D. D. Brown, R. R. Safford, I. H. Gessner, D. L. Gordon, and G. A. Ewy. 1999. Simulation technology for health care professional skills training and assessment. *JAMA*, 282(9):861–866.

Akinori Ito, Masaki Kohda, and Mari Ostendorf. 1999. A new metric for stochastic language model evaluation. In *Sixth European Conference on Speech Communication and Technology EUROSPEECH*.

R Iyer, Mari Ostendorf, and M Meteer. 1997. Analyzing and predicting language model improvements. In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 254–261.

Abigail Z Jacobs and Hanna Wallach. 2021. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 375–385, New York, NY, USA. Association for Computing Machinery.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3755–3763, Online. Association for Computational Linguistics.

Junmo Kang, Hongyin Luo, Yada Zhu, Jacob Hansen, James Glass, David Cox, Alan Ritter, Rogerio Feris, and Leonid Karlinsky. 2024. Self-specialization: Uncovering latent expertise within large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 2681–2706, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Colin Koopman. 2025. private communication.

Hur-Li Lee. 2000. What is a collection? *Journal of the American Society for Information Science*, 51(12):1106–1113.

Dongdong Li, Zhaochun Ren, Pengjie Ren, Zhumin Chen, Miao Fan, Jun Ma, and Maarten de Rijke. 2021. Semi-supervised variational reasoning for medical dialogue generation. In *SIGIR*.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. 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)*, pages 3469–3483, Online. Association for Computational Linguistics.

Wenge Liu, Jianheng Tang, Yi Cheng, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. 2022. MedDG: an entity-centric medical consultation dataset for entity-aware medical dialogue generation. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 447–459. Springer.

Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On LLMs-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11065–11082, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

William C. McGaghie, S. Barry Issenberg, Emil R. Petrusa, and Ross J. Scalese. 2010. A critical review of simulation-based medical education research: 2003-2009. *Medical Education*, 44(1):50–63.

Nikita Mehandru, Samantha Robertson, and Niloufar Salehi. 2022. Reliable and safe use of machine translation in medical settings. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 2016–2025, New York, NY, USA. Association for Computing Machinery.

Luise Modersohn, Stefan Schulz, Christina Lohr, and Udo Hahn. 2022. GRASCCO - The First Publicly Shareable, Multiply-Alienated German Clinical Text Corpus. *Studies in Health Technology and Informatics*, 296:66–72.

D. Moser, M. Bender, and M. Sariyar. 2024. Generating synthetic healthcare dialogues in emergency medicine using large language models. *Studies in Health Technology and Informatics*, 321:235–239.

Hajra Murtaza, Musharif Ahmed, Naurin Farooq Khan, Ghulam Murtaza, Saad Zafar, and Ambreen Bano. 2023. Synthetic data generation: State of the art in health care domain. *Computer Science Review*, 48:100546.

Bekelu Negash, Alan Katz, Christine J. Neilson, Moniruzzaman Moni, Marcello Nesca, Alexander Singer, and Jennifer E. Enns. 2023. De-identification of free text data containing personal health information: A scoping review of reviews. *International Journal of Population Data Science*, 8(1):2153.

Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Computational sociolinguistics: A Survey. *Computational Linguistics*, 42(3):537–593.

Tomohiro Nishiyama, Lisa Raithel, Roland Roller, Pierre Zweigenbaum, and Eiji Aramaki. 2024. Assessing authenticity and anonymity of synthetic user-generated content in the medical domain. In *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*, pages 8–17, St. Julian's, Malta. Association for Computational Linguistics.

Alex Papadopoulos Korfiatis, Francesco Moramarco, Radmila Sarac, and Aleksandar Savkov. 2022. 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)*, pages 588–598, Dublin, Ireland. Association for Computational Linguistics.

Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 39–48, Atlanta GA USA. ACM.

Samir Passi and Phoebe Sengers. 2020. Making data science systems work. *Big Data & Society*, 7(2):205395172093960.

Daniel S Quintana. 2020. A synthetic dataset primer for the biobehavioural sciences to promote reproducibility and hypothesis generation. *eLife*, 9:e53275.

J. P. Reiter. 2003. Inference for Partially Synthetic, Public Use Microdata Sets. *Survey Methodology*, 29(2):181–188.

R Röhrig et al. 2022. Grascco—the first publicly shareable, multiply-alienated german clinical text corpus. In *German Medical Data Sciences 2022-Future Medicine: More Precise, More Integrative, More Sustainable!: Proceedings of the Joint Conference of the 67th Annual Meeting of the German Association of Medical Informatics, Biometry, and Epidemiology EV (gmds) and the 14th Annual Meeting of the TMF-Technology, Methods, and Infrastructure for Networked Medical*, volume 296, page 66. IOS Press.

Donald B. Rubin. 1993. Discusion: Statistical Disclosure Limitation. *Journal of Official Statistics*, 9(2):461–468.

Tonya Rutherford-Hemming, Alaina Herrington, and Thye Peng Ngo. 2024. The Use of Standardized Patients to Teach Communication Skills—A Systematic Review. *Simulation in Healthcare: The Journal of the Society for Simulation in Healthcare*, 19(1S):S122–S128.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, Yokohama Japan. ACM.

Xiaoming Shi, Zeming Liu, Chuan Wang, Haitao Leng, Kui Xue, Xiaofan Zhang, and Shaoting Zhang. 2023. MidMed: Towards mixed-type dialogues for medical consultation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8145–8157, Toronto, Canada. Association for Computational Linguistics.

J. A. Starkweather, M. Kamp, and A. Monto. 1967. Psychiatric Interview Simulation by Computer. *Methods of Information in Medicine*, 06(01):15–23.

Aaron A. Tierney, Gregg Gayre, Brian Hoberman, Britt Mattern, Manuel Ballesca, Patricia Kipnis, Vincent Liu, and Kristine Lee. 2024. Ambient Artificial Intelligence Scribes to Alleviate the Burden of Clinical Documentation. *NEJM Catalyst*, 5(3).

U.S. Census Bureau. 2023. *U.S. Census Bureau Statistical Quality Standards*. U.S. Census Bureau, Washington DC, USA.

Ozlem Uzuner, Yuan Luo, and Peter Szolovits. 2007. Evaluating the state-of-the-art in automatic de-identification. *Journal of the American Medical Informatics Association: JAMIA*, 14(5):550–563.

Junda Wang, Zonghai Yao, Avijit Mitra, Samuel Osebe, Zhichao Yang, and Hong Yu. 2023a. UMASS_BioNLP at MEDIQA-chat 2023: Can LLMs generate high-quality synthetic note-oriented doctor-patient conversations? In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 460–471, Toronto, Canada. Association for Computational Linguistics.

Junda Wang, Zonghai Yao, Zhichao Yang, Huixue Zhou, Rumeng Li, Xun Wang, Yucheng Xu, and Hong Yu. 2024. NoteChat: A dataset of synthetic patient-physician conversations conditioned on clinical notes. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15183–15201, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Shanshan Wang, Yajing Yan, Rong Yan, Ting Li, Kaijie Ma, and Yani Yan. 2023b. A Chinese telemedicine-dialogue dataset annotated for named entities. *BMC medical informatics and decision making*, 23(1):264.

Kaishuai Xu, Yi Cheng, Wenjun Hou, Qiaoyu Tan, and Wenjie Li. 2024. Reasoning like a doctor: Improving medical dialogue systems via diagnostic reasoning process alignment. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6796–6814, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

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, Meliha Yetisgen, Jenny Huang, and Micah Grossman. 2020. Alignment annotation for clinic visit dialogue to clinical note sentence language generation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 413–421, Marseille, France. European Language Resources Association.

Stav Yosef, Moreah Zisquit, Ben Cohen, Anat Klomek Brunstein, Kfir Bar, and Doron Friedman. 2024. Assessing motivational interviewing sessions with ai-generated patient simulations. In *Proceedings of the 9th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2024)*, pages 1–11.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.

Hongbo Zhang, Junying Chen, Feng Jiang, Fei Yu, Zhihong Chen, Guiming Chen, Jianquan Li, Xiangbo Wu, Zhang Zhiyi, Qingying Xiao, Xiang Wan, Benyou Wang, and Haizhou Li. 2023. HuatuoGPT, towards taming language model to be a doctor. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10859–10885, Singapore. Association for Computational Linguistics.

Yuanzhe Zhang, Zhongtao Jiang, Tao Zhang, Shiwan Liu, Jiarun Cao, Kang Liu, Shengping Liu, and Jun Zhao. 2020. MIE: A medical information extractor towards medical dialogues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6460–6469, Online. Association for Computational Linguistics.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023. AugESC: Dialogue augmentation with large language models for emotional support conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568, Toronto, Canada. Association for Computational Linguistics.

A. What is a Dataset?

In response to the complexities surrounding the question of whether a given dataset is “synthetic” or not, we propose a conceptual framework whose aim is to provide structure and clarity when describing datasets that have a synthetic component. To facilitate meaningful comparisons, we must first define what we mean by “dataset,” and propose here a broad and “teleological” definition.We set as a foundation that a dataset is a finite collection of information-bearing objects, in the sense described by Buckland (1991). Such a collection may be either digital or physical in nature. Note that we refer here to a “collection” as opposed to a “grouping”,

"assembly", "set", etc. Not all groupings of information objects constitute datasets. A random and incidental set of bits would not constitute a dataset, nor would a random pile of books encountered on the roadside.

A *collection* implies a *collector* (i.e., an external agent with at least some degree of intentionality behind its assembly). Furthermore, a *collection* necessarily brings with it some notion of an intended *user* whose needs or goals the collection is meant to support.[12] Another key property of a *collection* is that it involves some kind of extrinsic criteria for what is and is not included in the collection.[13]

This definition, however, is insufficient. An intentionally collected set of bits representing random numbers may not *on its own* constitute a dataset. Therefore, we introduce a second constraint to our definition. Our collection of information-bearing objects must be *documents* in the sense meant by Briet (1951), who defined a document as "any concrete or symbolic indication, preserved or recorded, for reconstructing or for proving a phenomenon, whether physical or mental."[14]

According to this definition, a crucial distinction between "a collection of random bits" and a "dataset" is that the latter has an underlying *intention* of some sort on the part of whoever is identifying it as such (its creator or collector), and that this intention is reflected in the nature of the dataset's contents. The "concrete or symbolic indication" is being intentionally "preserved or recorded" *for* some purpose. Thus, we can conclude that it is necessary for that purpose that it be a *meaningful* and *representative* representation. Otherwise it would be useless for "reconstructing or ... proving a phenomenon."[15]

Under this part of our definition, we argue that part of what makes a dataset a dataset is that it is *intended* by its originators to be representative of *"something"*, hence, our definition's description as "teleological." What that "something" *is* may vary wildly, as may the degree to which a given dataset actually represents whatever it is meant to represent. These are important questions, but even a dataset that is flawed, incomplete, broken, etc., still counts as *a* dataset.

To illustrate this distinction, consider the bits representing the set of digital music files on one of the authors' laptops. This is comprised of a large grouping of binary files and associated metadata, but taken on its own, stripped of context and intention, it is simply that: a grouping. It serves no purpose and has no intrinsic meaning, other than to exist and be stored in his computer's filesystem. Its constituent bits may be interpretable by this or that piece of software, but are essentially the digital equivalent of a random and incidental pile of books.

In this particular case, it is reasonable to argue that they represent a *collection*, in the sense previously described. They have been intentionally collected, in that they have been purposefully assembled over a period of time as new music has been added, and there also exist numerous possible digital music files that are *not* in the collection. However, it would be nonsensical to say that this particular collection is or is not "representative", "meaningful", "useful", "complete", "well-formed", etc.: each of those descriptors implies a *telos* of some kind. In other words, each immediately prompts the questions, "for what?" and "for whom?" "Meaningful" — to whom, in what way, and in what context? The files in their present form lack any such thing — they are simply files.

Now, imagine that the author in question decides to try and *use* this collection as part of an analysis, perhaps to visualize his music consumption patterns. We argue that it is at this point, and not before, that our collection of information objects becomes a *dataset*: the point at which it is first intentionally collected *in order to represent, reconstruct, prove, etc. a particular phenomenon*. In other words, the point at which the bits become part of a process of problem formulation.[16]

Next, we add an additional constraint: in addition to the intention on the part of somebody to use a collection for some particular purpose, we also argue that there must also be an intention for some other party to understand the collection as being similarly meaningful. This is closely linked to the notion of "audience" in our discussion of collections above, but is more specifically linked to the *telos* of the dataset in question.

At this point, our definition has a "what" (collection of information objects), a "when" (the point at which somebody treats the collection as being a meaningful representation of a particular phenomenon), and a "who" (whatever additional audience or user community the author expects to

[12] The user and the collector may, of course, be the same entity.

[13] In the context of traditional library and archival collections, questions of ownership and accessibility also play into the definition of a collection (Lee, 2000).

[14] "Tout indice concret ou symbolique, conservé ou enregistré, aux fins de représenter ou de prouver un phénomène ou physique ou intellectual.", Briet (1951, p. 7) cited in Buckland (1991).

[15] In the case of a dataset comprising random noise, the intended point of the dataset is that the noise is in fact random according to some particular definition or scheme, and the existence of the dataset would constitute an assertion of that property on the part the dataset's authors.

[16] See Passi and Barocas (2019) for an excellent discussion on this topic, including on the contingent and bidirectional relationship between data and problem formulation.

align with them as to the nature and purpose of the dataset). In the interests of theoretical completeness, we add one more constraint, which may be categorized under the heading of "whether,"[17] and argue that for a collection to be considered a dataset, it must also contain information objects that in terms of their structure and content could plausibly be used from a functional or operational standpoint to achieve the dataset's purpose.

In other words, if the point of the dataset is to visualize the author's music consumption patterns, it must necessarily be comprised of documents containing the appropriate kinds of information (e.g., digital files containing music, metadata recording listening patterns, financial records of purchases, etc.), as opposed to being (e.g.) made up of random bits, pictures of cats, etc. This part of our definition may at first appear somewhat tautological. However, we include it to help rule out counterexamples and clarify the functional nature of our definition.[18] An analogy with measurement theory (Brewer and Crano, 2014) is appropriate. This part of our definition attempts to rhyme more with "face validity" than it does with "construct validity" or "measurement validity", in that we are not referring to how *well* the datasets' contents serve its intended purpose but rather are speaking about basic functional plausibility.

Nothing in our definition means that a dataset built to serve one *telos* may not be used in pursuit of another, or that a dataset may not be designed in pursuit of multiple *telea*. It does, however, help make apparent the fact that re-purposing of data is a less straightforward undertaking than it often appears and requires careful consideration. Additionally, our definition has nothing to say about what *kinds* of purposes are or are not "valid," or limits to how general or specific the purpose of a dataset may be.

B. Search Strategies

B.1. PubMed Search Strategy

Our PubMed search strategy consisted of three separate searches, with the goal of identifying published datasets of clinical dialogues, either synthetic or natural. This was conducted as a supplement to our preliminary literature search conducted using less precise methods, and should be considered current as of April, 2025.

The first search used solely keywords, and attempted to directly capture papers reporting the creation of datasets relating to dialogues, and was as follows: `(((patient AND provider) or (clinical)) AND (conversation* OR dialog OR dialogue)) AND (corpus OR dataset)`

The second attempted to leverage MeSH index terms, and did not specifically include terms relating to datasets or corpora, but did add a handful of keywords to reduce scope: `("Artificial Intelligence"[MeSH Terms] OR "Machine Learning"[MeSH Terms] OR "Natural Language Processing"[MeSH Terms] OR "Information Storage and Retrieval"[MeSH Terms] OR "Pattern Recognition, Automated"[MeSH Terms]) AND ("Physician-Patient Relations"[MeSH Terms] OR "Referral and Consultation"[MeSH Terms] OR "Interviews as Topic"[MeSH Terms] OR "Patient Simulation"[MeSH Terms] OR "Communication"[MeSH Terms]) AND ("conversation" OR "dialogue" OR "dialog" OR "discourse")`

The third returned to a keyword approach, and focused on patient-provider conversations: `(((patient AND provider) or (clinical)) AND (conversation* OR dialog OR dialogue)) AND (corpus OR dataset)`

The three searches together yielded 679 unique results, of which the vast majority were false positives. An LLM (Gemma 3 21b) was used to perform further filtration. First, we instructed the model to classify each retrieved paper according to whether its abstract described "automated processing of either spoken or written clinical language in the context of patient-provider communication." This reduced the set to 371 articles; manual inspection revealed satisfactory performance, and we were able to verify that specific known true-positive results were correctly classified by this method, further supporting its efficacy.

We, then, performed a second step of LLM-based filtration, instructing the model to classify each of those articles as to whether the abstract was "describing a paper that created a new corpus or dataset with the intention of it being re-usable by others." The model was instructed to prioritize recall over precision.

This reduced our set of articles to review to 125. Both prompts are rendered in the boxes from Appendix D. We used similar heuristic methods to verify that this filtering method performed at a satisfactory level of accuracy. The precision was reasonably high, in that articles in this set did in general contain notable discussion of the details of the datasets that they used, and generally did all describe the creation of novel corpora. However,

[17] In the descriptive sense of "whether or not it is possible", not the prescriptive sense of "whether or not it is a good idea".

[18] We express our thanks to Dr. Colin Koopman for this part of our definition (Koopman, 2025).

many were describing corpora that were outside of our review's scope (i.e., datasets of QA pairs, etc.), or described corpora that were not releasable due to reasons of privacy or other such constraints. Only 4 papers matched our criteria and were included in our review.

B.2. ACL Anthology Search Strategy

We implemented a script to retrieve, filter, and annotate papers from the ACL Anthology https://aclanthology.org/. The process involves scraping the HTML content of the full ACL proceedings between 2019-April, 2025 (EMNLP, ACL, EACL, NAACL, LREC) including all the workshops associated. Each paper title, abstract, authors and paper URL was extracted.

Papers were filtered based on the presence of the keyword `medic` in either the title or the abstract. This filtering is high recall, case-insensitive, and matches partial substrings, resulting a total of 591 papers. For each filtered paper, the corresponding full text document is extracted and submitted to the ChatPDF API.[19] A fixed prompt (see Appendix D) was used to request a structured JSON response with answers to the following binary and list-format questions: (1) whether a synthetic dataset was created, (2) whether the dataset contains dialogues, (3) whether human evaluation was conducted, (4) whether the dataset was used in a downstream task. We further reduce the list to 180 papers based on criteria (1) and (2). A manual review of the titles and abstracts further narrowed the set to 23 papers. Full-text examination of these 23 papers resulted in 11 that matched all criteria of interest and were included in our final review.

B.3. dblp Search Strategy

The search employed on https://dblp.org was broad and covered a range of related fields, including speech processing, medical informatics, dialogue systems, artificial intelligence, argumentation, and information retrieval. To ensure high recall, [20] the search targeted titles containing words beginning with `med*` (e.g., medical, medication), `syn*` (e.g., synthesis, synthetic), or `sim*` (e.g., simulation, simulating) in combination with `dialo*` (e.g., dialogue).

To refine the results, a filter was applied to exclude publications from major venues associated with the Association for Computational Linguistics (ACL) and preprints. The final corpus was restricted to articles published in selected conferences, workshops, and journal series,[21] including *INTERSPEECH*, *EUROSPEECH*, *ICASSP*, *IJCAI*, *SIGIR*, *AAAI*, *MEDINFO*, the *Journal of Biomedical Informatics*, *BMC Medical Informatics and Decision Making*, *Computers in Biology and Medicine*, *CEUR-WS*, *FAIA*, *CCIS*, *SCI*, and relevant Springer proceedings, among others.

In total, 356 papers were identified through this search. Based on title screening, 45 of the retrieved papers were selected for further manual review of abstract and content. Among these, only 10 papers have been marked as suitable for our review.

[19] https://www.chatpdf.com/

[20] `(med* | syn* | sim*) & dial*`

[21] In total 362 `streamid's` corresponding to different publications.

C. Synthetic Datasets

Source	Initial Hits	After First Filter	After Review	Included
PubMed	679	371 (LLM: clinical language)	125 (LLM: dataset creation)	4
ACL Anthology	591	180 (LLM: dataset+dialogue)	23	11
dblp	356	-	45	10

Table 2: Search strategy summary. The first filter for PubMed and ACL Anthology are based on LLM prompting. There is an overlap between the three search strategies and the total number of unique papers is 20.

Dataset	Acronym	Language	Types
Ben Abacha et al. (2023)	MTS-DIALOG	en	Human Type 3, Machine Type 1
Liu et al. (2021)	ESConv	en	Human Type 3, Machine Type 1
Fareez et al. (2022)		en	Human Type 3, Machine Type 2
Papadopoulos Korfiatis et al. (2022)	PriMock57	en	Human Type 3, Machine Type 2
Yim et al. (2023)	ACI-bench	en	Human Type 3, Machine Type 2
ALMutairi et al. (2024)		ar-sa	Human Type 3, Machine Type 3
Zheng et al. (2023)	AugESC	en	Human Type 3, Machine Type 3
Chen et al. (2023b)	IMCS-21	zh	Human Type 2, Machine Type 1
Shi et al. (2023)	MidMed	zh	Human Type 2, Machine Type 2
Zeng et al. (2020)	MedDialog	zh / en	Human Type 1, Machine Type 1
Liu et al. (2022)	MedDG	zh	Human Type 1, Machine Type 2
Li et al. (2021)	KaMed	zh	Human Type 1,Machine Type 2
Xu et al. (2024)	EMULATION	zh	Human Type 1, Machine Type 2
Binici et al. (2025)	MEDSAGE	ch	Human Type 1, Machine Type 2
Zhang et al. (2023)	HuatuoGPT	zh	Human Type 1, Machine Type 3
Nishiyama et al. (2024)		ja	Human Type 1, Machine Type 3
Han et al. (2024)	PSYDIAL	ko	Human Type 1, Machine Type 3
Chen et al. (2023a)	DoPaCo	en	Human Type 1, Machine Type 3
Wang et al. (2024)	NoteChat	en	Human Type 1, Machine Type 3
Moser et al. (2024)		en, de	Human Type 1, Machine Type 3

Table 3: The list of identified synthetic datasets and their categories.

D. LLM Prompts

Gemma 3 27b: clinical language filter

Your role is to review an abstract of a journal article, and decide whether it should be included in a literature review. We are looking for articles about automated processing of either spoken or written clinical language in the context of patient-provider communication. Examples would include: clinical conversations, discourse, written narrative communications, etc. Either transcribed or recorded language is acceptable.
If you are not sure, or if the article is borderline, please consider it in-bounds and err on the side of inclusion.

Gemma 3 27b: dataset creation filter

Your role is to review an abstract of a journal article, and determine whether or not the abstract is describing a paper that created a new corpus or dataset with the intention of it being re-usable by others. Possible things to look for:
- The abstract might explicitly mention that the dataset produced as part of the article was released or otherwise made publicly available.
- The abstract might talk about how the authors developed a novel corpus, specifically because of a lack of already-existing data or because existing data was insufficient in some way.
Remember, all of these papers will have involved making a dataset or corpus of some kind. We are looking specifically for papers that where one of the final products of the research, or where the main point of the research, was to create a new dataset or corpus, especially if it is described as being released or made available.
Simply mentioning an already-existing dataset is not enough.

ChatPDF: dataset+dialogue

Provide a json response using the following structure:

```
{
"synthetic_dataset_created": true/false,
"contains_dialogues": true/false,
"human_evaluation": true/false,
"uses_dataset_in_downstream_task": true/false
}
```

To answer the following questions:
- Do the authors create a synthetic dataset in this paper?
- Does it contain dialogues?
- Does it have a human evaluation of the dataset?
- Does it use the dataset in a downstream task?