Synthetic Biomedical Data Generation Via a Beam Tree Strategy for Large Language Models

Juan Cano-Benito[1], Sven Hertling[2], Fabian Berns[3] and Heiko Paulheim[2]

[1]*Universidad Politécnica de Madrid, Madrid, Spain*
[2]*Mannheim University, Mannheim, Baden-Württemberg, Germany*
[3]*medicalvalues GmbH, Karlsruhe, Baden-Württemberg, Germany*

Abstract

In recent years, the generation of synthetic data in the biomedical domain has become increasingly important in addressing the limited availability of public datasets and the imbalance that exists in certain datasets. However, generative models often present problems, such as repetitive phrases and hallucinations, that limit the quality and reliability of the data produced. This paper proposes a method for creating synthetic biomedical sentences using a combination of beam tree search and knowledge graphs. The approach explores the generation of datasets in order to maximise the diversity of the phrases generated, while the incorporation of knowledge graphs guides the identification of valid biomedical entities and balances their distribution. The method is evaluated using three benchmark datasets (BC5CDR, ChemProt, and MedMentions) to measure its efficiency in generating synthetic data according to each dataset.

Keywords

Large Language Model, Synthetic dataset generation, Biomedical domain, Knowledge graph

1. Introduction

Over the past decade, the use of AI to organise biomedical knowledge has increased, from clinical text recognition to decision-making in hospitals [1, 2]. In recent years, industry and academia have analysed the potential of generative AI based on transformer models in the medical domain [3]. Generative AI shows potential in reasoning capabilities with few examples and in creating large amounts of domain-specific data on demand, with the potential to augment existing biomedical datasets, which are often limited in size and coverage [3].

Currently, one of the main challenges in the biomedical field is the lack of publicly available datasets [4]. As a result, most public datasets are based on scientific publications which, despite their value, tend to cover a limited range of biomedical entity types and relationships. These datasets often cover only a small number of entities. Broader datasets exist, but they are often highly unbalanced [5].

To solve the previously mentioned issues, models can be used to create synthetic data. The creation of these datasets has certain challenges. For example, generative models tend to show a fixation on entities, i.e., a bias toward repeatedly producing the same diseases, chemicals, or high-frequency conditions, making it difficult to obtain a balanced corpus with diverse and rarely mentioned entities [6]. Furthermore, the hallucinations characteristic of generative models are a significant problem, as generative models can invent non-existent drugs, impossible biological interactions, or clinically unsafe associations, creating a problem in the implementation of synthetic data generation and classification systems in the biomedical domain [7].

Although the generation of synthetic data using generative AI has already been explored [8, 9, 10, 11, 12, 13, 14, 15], many existing approaches reduce the generation process to prompts or filters, limiting control over entity diversity, contextual consistency, and domain alignment. These methods often lack mechanisms to prevent the model from generating similar repetitive results.

SeWebMeDA-2026: 9th International Workshop on Semantic Web Solutions for Large-scale Biomedical Data Analytics, May 10, 2026, Dubrovnik, Croatia
juan.cano@upm.es (J. Cano-Benito); sven.hertling@uni-mannheim.de (S. Hertling); fabian.berns@medicalvalues.de (F. Berns); heiko.paulheim@uni-mannheim.de (H. Paulheim)
0000-0002-5638-4977 (J. Cano-Benito); 0000-0003-0333-5888 (S. Hertling); 0000-0003-4386-8195 (H. Paulheim)

In this article, we present an approach that combines a beam tree search strategy with semantic filtering based on knowledge graphs. The proposed method uses a new generation strategy base on beam search to explore the space of candidate sentences, while using domain-specific entities and ontological information to guide the model in generating entities. By integrating dataset construction using beam tree search or knowledge-based constraints, our approach aims to produce synthetic biomedical sentences that are diverse, balanced, and semantically aligned with medically grounded concepts.

The article is organised as follows: Section 2 presents the related work; Section 3 proposes a method to generate synthetic datasets using large language models with a beam search strategy; Section 4 explains the generation process; Section 4 evaluates the results; Section 5 discusses the results obtained, and finally section 6 presents the conclusions of this paper and future work.

2. Related Work

A Large Language Model (LLM) is a neural network trained on massive structured data (text corpora, database tables, RDF, etc) using transformer architectures, enabling text generation for specific tasks. Thus, it allows the expansion or creation of new datasets that are typically costly to obtain or extend [16]. In general, in the generation of synthetic datasets there are remaining challenges: LLMs can introduce factual inaccuracies or biases, producing false and unnatural data [16]. Therefore, recent work emphasises the need for quality control when generating datasets. Strategies such as limiting generation to contexts known to the model, iteratively refining outputs, and applying filters to discard questionable examples have been proposed [16].

For example, in natural language processing, research has explored the use of models to generate labelled text (e.g., sentences with their category) to train classifiers, reducing the need for manual annotation [17]. However, the effectiveness of synthetic data generated by LLMs may vary depending on the task. Li et al. [17] observed models trained with LLM data performed significantly worse when using synthetic data for text classification in tasks with a higher subjective component.

The biomedical field has adopted these techniques to mitigate the lack of publicly available clinical data and protect patient privacy. The most studied strategy in recent years has been prompt-based generation, in which models such as GPT-4 are used to directly generate synthetic clinical data [9, 11, 10, 13, 14, 15]. This includes the generation of synthetic datasets [9, 10], the generation of clinical notes [11, 12], question-answer pairs [13], and synthetic dialogues between doctors and patients [14, 15]. Some efforts have complemented this with filtering and quality control, often using heuristic rules or another LLM to assess plausibility, naturalness, or veracity [12, 15]. A recent line of work generates synthetic data providing a prompt to the model and rely on the resulting output [8, 11, 12]. Other approaches employ tree search strategies (similar to beam search) to explore multiple decoding paths [18, 19]. However, this tree search method has not been explored in the medical domain.

Before using medical synthetic data in real scenarios, it is important to evaluate the quality of the data to ensure its value [9, 20, 21]. The main current evaluation methods can be divided into two categories: direct, which assesses the intrinsic quality of the synthetic data itself (e.g., fidelity and diversity), and indirect, which assesses the data through their impact on the performance of the subsequent model, including benchmarking and open evaluation [9]. In automatic evaluation, the main metrics can be summarised in four: precision (exact match, F1 score, and ROUGE score), calibrations (expected calibration error and area under the curve), fairness (demographic parity difference and equalised odds diference) and robustness (attack success rate and performance drop rate) [22]

Other proposals [21] adopt a more technical and granular approach, focussing on statistical tests, anomaly modelling, privacy attacks, and usability. In contrast, a specific framework for evaluating synthetic data in biomedicine is the Synthetic Medical Data Scorecard (SMDS) [20], which proposes to evaluate generated datasets using seven dimensions: consistency with real data, coverage of variability, compliance with clinical constraints, completeness, privacy and standards, clarity of generation, and consistency between subgroups. This approach offers a more structured and clinical guide than classic

fidelity and/or diversity metrics, allowing errors to be detected (e.g. anatomically incorrect entities or clinical hallucinations). These criteria are flexible and should be adapted to each type of synthetic dataset.

3. Method

This section describes the method used for the generation of synthetic biomedical data, using three different datasets as a reference: BC5CDR (a dataset that covers diseases and chemicals), BioRED (covering six types of biomedical entities) and MedMentions (a dataset covering 21 UMLS semantic types). Our method uses an indication construction process guided by a knowledge graph and a beam-search generation strategy. Due to the sensitive nature of biomedical information and in order to reuse this method, large open language models (in this case, OSS-GPT20B) have been used for all experiments, although the method is applicable to all LLMs. The diagram of this method is depicted in Figure 1.

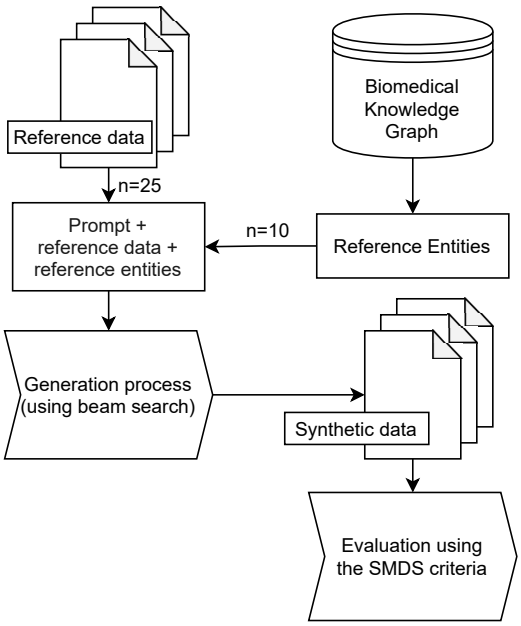

Figure 1: Overview of the synthetic data generation process.

First, the generation process begins by sending to the model a prompt, the prompt receives samples of sentences from the original dataset and samples of biomedical entities stored in a knowledge graph. These examples and entities serve to guide the model in determining how sentence generation should be structured. The model is instructed not to copy these entities directly but to use them as examples.

Instead of relying on a single deterministic generation path, our method performs an exploration of multiple candidate tokens. A tree-structured branching mechanism is used that expands several candidate prefixes in the early stages of generation, when lexical and structural decisions have a major impact. In our method, using a pure beam search almost always resulted in generating the same entities. Therefore, a structure similar to that of the beam search is used. Instead of always selecting the most probable tokens at each step, we introduce stochastic sampling to choose among the most relevant candidates.

This allows different temperatures to be applied during generation, causing the model to choose safer options (lower temperatures) or explore less frequent alternatives (higher temperatures). Although the procedure shares with beam search the idea of simultaneously exploring different paths, it does not behave like a pure beam search because it does not preserve the best paths deterministically, but allows controlled deviations through temperature. As the sentence grows, branching decreases as fewer variations are required to maintain diversity, and subsequent tokens tend to rely more on the

preceding context, increasing lexical and structural diversity. To determine the depth of the token search, behaviour is modelled using a power law scaling function:

$$f(x) = a \cdot x^b \tag{1}$$

Where a is a scaling coefficient and b is the decay exponent. If $a = 500$ and $b = -4$ are set, the resulting curve decreases as x increases. Since the function yields non-integer values that can approach zero at higher indices, the function is discretised by applying the floor operator and applying a lower bound of 1 to avoid numerical collapse. Formally, the discrete sequence $S = \{s_i\}$ is defined as:

$$s_i = \max\left(1, \left\lfloor a \cdot i^b \right\rfloor\right), \quad i \in [1, n) \tag{2}$$

This transformation ensures that the initial values reflect the steep decline predicted by the power-law behaviour, while later elements are stabilise at 1 due to the negative exponent. Figure 2 depicted the number of beams per token.

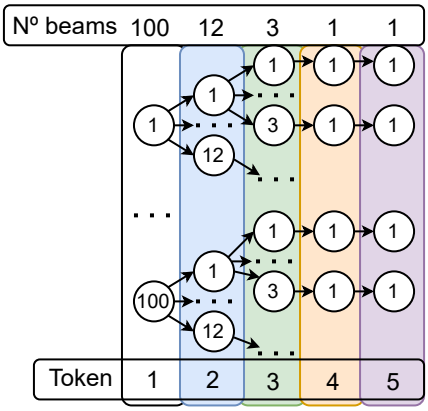

Figure 2: Number of beams per token applying the power law decay.

For each entity in each dataset, to have a balanced results between domains, $a = 100$ and $b = -3$ have been used in BC5CDR with 2 domains, generating a total of 3,600 sentences, $a = 114$ and $b = -3$ in BioRed with 6 domains, generating a total of 6,384 sentences, and $a = 121$ and $b = -3$ in MedMentions with 21 domains, generating a total of 7,260 sentences. To perform controlled generation, a percentage of the entities to be generated is chosen (in this case, a balanced sample has been chosen, so that if the number of beams is 100 and the dataset is BC5CDR, which contains 2 entities, 50 and 50 will be generated). In addition, different temperatures have been used in the generation of synthetic sentences. All code is public available on GitHub[1].

4. Experimental results

In this section, the results are evaluated using the SMDS [20], once the data have been generated, which consists of seven criteria. Therefore, each criterion and the methods used in each one are detailed.

Congruence (1C). This criterion measures how closely the statistical properties of the synthetic dataset match those of the real dataset. Three embedding based similarity metrics included in the SMDS are evaluated, where each individual sentence is encoded:

- *Cosine similarity (mean)*: the cosine similarity between the mean embeddings of real and synthetic data.
- *JS divergence*: Jensen-Shannon divergence between distributions of pairwise cosine distances.

[1]https://github.com/jucanbe/SyntheticBiomedicalData-BeamSearch

- *Earth Mover's Distance*: implemented as the Wasserstein distance between distributions of pairwise distances.

Coverage (2C). This criterion evaluates the extent to which the synthetic dataset covers the semantic space of the real corpus. The SMDS includes the three embedding-based metrics adapted for text:

- *Recall (%)*: vocabulary recall between real and synthetic corpora.
- *Convex hull ratio*: ratio between synthetic and real convex hull areas.
- *Coverage (%)*: proportion of sentences in the real dataset that have at least one similar counterpart in the synthetic dataset (in our case, a sentence with embedding similarity ≥ 0.8).

Constraint (3C). This criterion verifies whether synthetic sentences satisfy basic structural constraints.

- *Constraint violation rate (%)*: percentage (constraint violation rate $> 0.6\%$) of synthetic sentences that violate some structural requirements (is not empty, contains a period, and is a single sentence).
- *Constraint-valid samples (%)*: complement of the above, also reported in the SMDS as a standard form.

Completeness (4C). This criterion evaluates whether each synthetic sample contains the biomedical content required.

- *Required clinical field (%)*: percentage of synthetic sentences containing at least one recognised biomedical entity.
- *Missing data (%)*: percentage of sentences lacking any such entity.

Compliance (5C). This criterion refers to preserving privacy and avoiding memorisation. Since metrics based on data protection and anonymisation of the SMDS are not applicable to free text, two standard indicators of textual privacy are applied:

- *Nearest-real cosine similarity (mean)*: measures the proximity of synthetic sentences to their closest actual neighbors in the embedding space, which serves as an indicator of re-identification risk.
- *Suspected memorised samples*: proportion of synthetic sentences whose closest real neighbors exceed both the cosine similarity and the Jaccard threshold of n-grams.

Comprehension (6C). This criterion refers to the quality of the documentation and the interpretability of the synthetic data generation process. Since it is not a quantifiable numerical metric, it is omitted from the results, but this criterion is met with the paper and the GitHub repo.

Consistency (7C). This criterion measures the stability across subgroups of the synthetic data. Following the SMDS, we compute:

- *Variance of domain*: variance of cosine similarities at the domain level with respect to the global embedding mean.
- *Max-min difference*: difference between the highest and lowest similarity scores at the domain level.

4.1. BC5CDR

In this subsection, the synthetic dataset based on BC5CDR is evaluated (3,600 synthetic sentences). Table 1 shows the results obtained applying the SMDS. This dataset contains 2 domains: disease (a pathological or medical condition) and chemical (Drug, compound, molecule, or medication)

Congruence (1C). Across all temperatures, the cosine similarity between the mean embeddings of real and synthetic corpora remains high (> 0.993), indicating that the global semantic structure of the

Metric		T=0.8	T=0.6	T=0.4	T=0.2	T=0.01
1C	Cosine similarity (mean)	0.9939	0.9941	0.9958	0.9951	0.9957
	JS divergence (pairwise dist.)	0.0147	0.0286	0.0056	0.0085	0.0053
	Earth Mover's distance (Wasserstein)	0.0158	0.0317	0.0071	0.0120	0.0090
2C	Recall (%)	58.17	58.49	55.42	56.57	53.78
	Convex hull ratio	0.5902	0.7052	0.5626	0.6294	0.4891
	Coverage (%)	99.86	99.94	99.92	99.92	99.90
3C	Constraint violation rate (%)	0.8333	0.6944	0.1667	0.0833	0.0556
	Constraint-valid samples (%)	99.17	99.31	99.83	99.92	99.94
4C	Required clinical field (%)	99.64	99.64	99.64	99.83	99.58
	Missing data (%)	0.36	0.36	0.36	0.17	0.42
5C	Nearest-real cosine similarity (mean)	0.9865	0.9847	0.9870	0.9861	0.9863
7C	Variance of domain	1.28×10^{-12}	3.27×10^{-11}	2.27×10^{-13}	3.09×10^{-12}	8.53×10^{-12}
	Max-min difference	2.26×10^{-6}	1.14×10^{-5}	9.54×10^{-7}	3.51×10^{-6}	5.84×10^{-6}

Table 1
Evaluation of synthetic BC5CDR data following the SMDS.

dataset is well preserved. Lower temperatures (T=0.2 and T=0.01) produce the highest cosine similarity overall.

Both JS divergence and Earth Mover's Distance increase as the temperature increases, meaning that higher temperatures introduce more variability in local distance distributions. The best congruence is observed at **T=0.01**, which achieves the lowest JS divergence (0.0053) and the lowest Wasserstein distance (0.0063).

Coverage (2C). Vocabulary recall steadily decreases as the temperature decreases, meaning that higher temperatures generate a broader lexical range. The convex hull ratio shows moderate variation, with T=0.2 and T=0.01 producing the tightest alignment, while T=0.6 yields the widest synthetic spread.

The coverage is $\approx$ 100% in all temperatures, confirming that nearly every real sentence has a very close synthetic neighbour, regardless of the generation temperature.

Constraint (3C). The satisfaction of constraints varies depending on the temperature. Higher temperatures (T=0.8 and T=0.6) produce more structural errors, while very low temperatures (T=0.2 and T=0.01) reduce the violation rate.

The most accurate sentences in terms of structure are produced at **T=0.01**, achieving the lowest violation rate (0.0026) and the highest proportion of valid samples in terms of constraints (99.74%).

Completeness (4C). The BC5CDR synthetic sentences contain a large number of biomedical entities. The proportion of sentences containing at least one entity exceeds 99.6% at all temperatures, with missing data remaining below 0.4%. Although differences are minimal, T=0.8 and T=0.6 achieve the highest completeness (99.64–99.66%).

Compliance (5C). The closest real cosine similarity remains stable across all temperatures (0.984–0.987), indicating a low risk of re-identification. No memorized phrases were detected at any temperature, meaning that the LLM does not reproduce content almost verbatim from the training set.

Consistency (7C). Since BC5CDR contains only two domains with very homogeneous embedding distributions, both the variance of the score at the domain level and the maximum–minimum difference are practically zero at all temperatures. This indicates stability and uniform behaviour across all chemical and diseases synthetic sentences.

Analysing the synthetic dataset. Taking into account all criteria, the balance dataset is obtained

at **T=0.01**. This temperature has a good semantic congruence (best cosine similarity and lowest JS divergence and Wasserstein distance), lower restriction violation rate, high completeness, no signs of memorisation, and good domain consistency.

4.2. Biored

In this subsection, the synthetic dataset based on biored is evaluated (6,384 synthetic sentences). Table 2 shows the results obtained applying the SMDS. This data set contains 6 domains: CellLine (a specific cell line used in biomedical research), ChemicalEntity (a chemical compound, drug or small molecule), DiseaseOrPhenotypicFeature (a disease or observable trait), GeneOrGeneProduct (a gene or its expressed product), OrganismTaxom (a species or strain) and SequenceVariant (a specific variation in a DNA/RNA/protein sequence).

	Metric	T=0.8	T=0.6	T=0.4	T=0.2	T=0.01
1C	Cosine similarity (mean)	0.9952	0.9942	0.9939	0.9935	0.9939
	JS divergence (pairwise dist.)	0.0682	0.0757	0.0790	0.0923	0.1006
	Earth Mover's distance (Wasserstein)	0.0581	0.0631	0.0704	0.0740	0.0863
2C	Recall (%)	63.83	61.14	59.68	58.68	59.56
	Convex hull ratio	0.7964	0.7235	0.8458	0.8415	0.9853
	Coverage (%)	99.98	99.98	99.98	99.98	99.97
3C	Constraint violation rate (%)	26.77	32.59	29.70	28.10	26.53
	Constraint-valid samples (%)	73.23	67.40	70.30	71.90	73.46
4C	Required clinical field (%)	99.79	99.87	99.76	99.72	99.69
	Missing data (%)	0.20	0.13	0.24	0.28	0.31
5C	Nearest-real cosine similarity (mean)	0.9800	0.9800	0.9790	0.9800	0.9798
7C	Variance of domain	0.0002	0.0004	0.0004	0.0004	0.0004
	Max-min difference	0.0392	0.0534	0.0518	0.0536	0.0540

Table 2
Evaluation of synthetic Biored data following the SMDS.

Congruence (1C). The cosine similarity between the real and synthetic embedding means remains also high at all temperatures (> 0.993), indicating that the structure of BioRED is preserved. JS divergence and Earth Mover's Distance increase with temperature, as expected. Higher temperatures introduce more variability in the distance distribution. **T=0.01** has a more realistic behaviour, giving the lowest JS divergence (0.1006) and the lowest Earth Mover's Distance (0.0863).

Coverage (2C). Vocabulary recall decreases as temperature decreases, and T=0.8 reaches the highest recall (63.83%). This suggests that higher temperatures generate more diverse lexical content. The convex hull ratio shows fluctuations, indicating that the shape of the embedding space varies with the generation temperature. However,, coverage based on nearest neighbour similarity remains high across all temperatures ($\approx 100\%$).

Constraint (3C). Constraint satisfaction shows a dependence on temperature. Higher temperatures (T=0.8 and T=0.6) yield a high violation rate (> 26%), showing that increased randomness leads to more structural inconsistencies in the generated text. Lower temperatures improve structure, with the best results at **T=0.01**, which reduces the violation rate to 26.53% (still high compared to BC5CDR, but expected for BioRED due to its greater complexity).

Completeness (4C). Completeness remains good at all temperatures. The proportion of sentences that contain at least one biomedical entity remains above 99.7%, and missing data never exceeds 0.31%. This indicates that the generator includes domain-relevant information, even when the text becomes more random at higher temperatures.

Compliance (5C). Cosine similarity remains stable at all temperatures (0.979–0.980), indicating a low risk of re-identification. No memorised samples were detected at any temperature.

Consistency (7C). Variance and maximum-minimum difference across the six domains remain small for all temperatures. The generator does not favor specific types of entities or domains.

Analysing the synthetic dataset. The best synthetic dataset is **T=0.01**. This temperature has the lowest JS divergence and Wasserstein distance, good coverage, the highest valid proportion of constraints among the tested configurations, high completeness with minimal missing data, and consistent behaviour across all domains.

4.3. MedMentions

In this subsection, the synthetic dataset based on MedMentions is evaluated (7262 synthetic sentences). Table 3 shows the results obtained applying the SMDS. This dataset contains 21 domains: AnatomicalStructure (specific parts of the body or anatomical regions), Bacterium (bacterial organisms mentioned in the text), BiologicFunction (biological or physiological processes and functions), BiomedicalOccupationOrDiscipline (biomedical roles or fields of specialization), BodySubstance (substances originating from the body), BodySystem (functional body systems), Chemical (chemical substances and compounds), ClinicalAttribute (clinical characteristics or descriptors), Eukaryote (eukaryotic organisms such as parasites or fungi), Finding (clinical findings or observations), Food (food items or nutrients), HealthcareActivity (healthcare-related activities that are not procedures), InjuryOrPoisoning (injuries, poisonings, and related conditions), IntellectualProduct (guidelines, questionnaires, and other intellectual artifacts), MedicalDevice (medical instruments or devices), Organization (institutions or organizations), PopulationGroup (groups of people or patient populations), ProfessionalOrOccupationalGroup (professional categories or work groups), ResearchActivity (research-related activities or study types), SpatialConcept (spatial or locational concepts relevant to medicine), and Virus (viral agents).

Congruence (1C). As in the previous datasets, the cosine similarity remains high at all temperatures (> 0.987). JS divergence and Earth Mover's Distance follow the expected behaviour: higher temperature, greater divergence. The lowest temperature (T=0.01) produces the lowest JS divergence (0.0432) and the lowest Earth Mover's Distance (0.0287).

Coverage (2C). Vocabulary recall decreases steadily with temperature, with T=0.8 providing the widest lexical range (37.54%) and T=0.01 producing the lowest recall (30.83%). This is due to the fact that MedMentions covers a broad set, so lower temperatures tend to produce more confident and less diverse sentences. The convex hull ratio varies between the different temperatures, showing the complexity of the embedding distribution. Coverage remains high ($\approx$ 99.7%), which means that nearly all real sentences have a synthetic neighbour regardless of temperature.

Constraint (3C). The violation of restrictions decreases with temperature. Higher temperatures generate more structural irregularities, with T=0.8 reaching a violation rate of 0.3994%. The best behaviour is observed for T=0.01, where the violation rate drops to 0.1545% and the proportion of valid sentences reaches 99.85%. Compared to BioRED, MedMentions shows more stability at different temperatures, probably because its syntactic structures are more descriptive than relational.

Completeness (4C). The synthetic MedMentions dataset shows that each synthetic sentence contains at least one recognised biomedical entity (100%). This behaviour is expected given the structure of the dataset. MedMentions includes 21 semantic domains with high lexical diversity, which increases the

Metric		T=0.8	T=0.6	T=0.4	T=0.2	T=0.01
1C	Cosine similarity (mean)	0.9882	0.9882	0.9873	0.9886	0.9881
	JS divergence (pairwise dist.)	0.0491	0.0571	0.0445	0.0381	0.0432
	Earth Mover's distance (Wasserstein)	0.0339	0.0332	0.0277	0.0278	0.0302
2C	Recall (%)	37.54	34.34	33.82	30.68	30.83
	Convex hull ratio	0.5602	0.6573	0.4841	0.5046	0.5347
	Coverage (%)	99.86	99.80	99.66	99.78	99.78
3C	Constraint violation rate (%)	0.3994	0.3581	0.2342	0.3581	0.4545
	Constraint-valid samples (%)	99.60	99.64	99.76	99.64	99.54
4C	Required clinical field (%)	100	100	100	100	100
	Missing data (%)	0	0	0	0	0
5C	Nearest-real cosine similarity (mean)	0.9580	0.9572	0.9600	0.9593	0.9589
7C	Variance of domain	1.78×10^{-5}	1.75×10^{-5}	1.14×10^{-5}	1.35×10^{-5}	2.43×10^{-5}
	Max-min difference	0.0186	0.0187	0.0148	0.0164	0.0227

Table 3
Evaluation of synthetic MedMentions data following the SMDS.

probability that any generated sentence will match at least one entity from the real dataset. Furthermore, the prompting strategy leads the model to include medical entities, which reinforces the appearance of the entity regardless of the temperature. However, the metric captures the presence of domain relevant terminology, but does not reflect whether the distribution of entities or contextual usage matches the actual dataset.

Compliance (5C). The closest real cosine similarity remains stable (0.958–0.960) at all temperatures, indicating a low risk of reidentification and no evidence of memorisation. As in the previous synthetic datasets, the model does not generate the content verbatim.

Consistency (7C). Both the variance between domains and the maximum-minimum difference remain low for all temperature settings. Despite the large number of domains (21), the quality of the synthetic data remains stable between different types of entities.

Analysing the synthetic dataset. The best synthetic dataset is obtained with **T=0.01**. It exhibits the best or near-best congruence (lowest JS divergence), lowest structural error rate, better completeness, good domain-level consistency, and stable performance without memorisation.

5. Discussion

Table 4 summarises the synthetic datasets with the best results obtained for each dataset (BC5CDR, BioRED, and MedMentions), that after analysing each synthetic dataset, is T=0.1 in all dataset. Although the three corpora differ in size, heterogeneity, and semantic composition, the SMDS results show that these characteristics of the datasets have an impact on the quality of the synthetic data produced. In general, lower temperatures produce more stable and predictable texts that better match the structure of real datasets. However, the results are not good for all the datasets. Although BC5CDR and BioRED achieve high fidelity, MedMentions shows the limitations of current LLM-based synthetic generation when applied to a highly heterogeneous biomedical dataset.

Metric		BC5CDR	BioRed	MedMentions
1C	Cosine similarity (mean)	0.9957	0.9939	0.9881
	JS divergence (pairwise dist.)	0.0053	0.1006	0.0432
	Earth Mover's distance (Wasserstein)	0.0090	0.0863	0.0302
2C	Recall (%)	53.78	59.56	30.83
	Convex hull ratio	0.4891	0.9853	0.5347
	Coverage (%)	99.90	99.97	99.78
3C	Constraint violation rate (%)	0.0556	26.53	0.4545
	Constraint-valid samples (%)	99.94	73.46	99.54
4C	Required clinical field (%)	99.58	99.69	100
	Missing data (%)	0.42	0.31	0
5C	Nearest-real cosine similarity (mean)	0.9863	0.9798	0.9589
7C	Variance of domain	8.53×10^{-12}	0.0004	2.43×10^{-5}
	Max-min difference	0	0.0540	0.0227

Table 4
Results with T=0.01 in each synthetic dataset.

5.1. BC5CDR.

BC5CDR is the smallest and least diverse corpus, as it only contains two domains (diseases and chemistry). As expected, this results in better performance when generating synthetic data. The congruence metrics achieve very good values (cosine similarity 0.9957, JS divergence 0.0053, EMD 0.0090), reflecting close alignment with the actual embedding space. Coverage is almotst perfect (99.90%), although vocabulary recall remains moderate (53.78%), suggesting reliance on high-frequency terminology. Constraint violations are minimal (0.0556%), indicating that the model has less difficulty generating valid sentences for this domain. Completeness is high (99.58%), and consistency across the two domains is near perfect (variance = 0). It can be concluded that BC5CDR yields the best results for generating synthetic biomedical text with this method, as it is a dataset focused on its main entities.

5.2. BioRED.

BioRED is a dataset with six domains and more complex biomedical relationships, making it more challenging than BC5CDR. Despite this increase in complexity, the results are good. Congruence is still high (cosine similarity 0.9939), but divergence metrics (JS = 0.1006, EMD = 0.0863) indicate a wider spread in the embedding space. BioRED obtains the highest vocabulary recall (59.56%), which implies a richer lexical diversity than BC5CDR. However, the constraint violation rate increases to 26.53%, which means that more than one in four synthetic sentences fails to meet the structural requirements. Completeness remains high (99.69%), but the increased variability across domains is reflected in non-zero consistency

metrics. Therefore, the synthetic BioRED dataset offers a useful middle ground: synthetically generated text remains semantically diverse and more or less aligned with the real data, but structural correctness becomes a challenge.

5.3. MedMentions.

MedMentions is the most difficult corpus for synthetic data generation. It contains 21 domains covering anatomy, chemicals, organisms, clinical findings, research activities, procedures, etc., resulting in a broad and uneven semantic space. This complexity has an impact on the synthetic quality. Congruence metrics decrease compared to the other datasets (cosine similarity 0.9881), and while divergence values remain stable relative to BioRED, these values are still higher than in BC5CDR. Vocabulary recall drops to 30.83%, which shows the inability of the model to reproduce the lexical range of the real corpus. The constraint violation rate increases to 45.45%, which is the worst of all datasets (nearly half of the generated sentences violate structural rules). Although completeness reaches 100% %, this result is explained by the large number of entities (almost any biomedical term is considered a recognised entity) and not by the fidelity of the dataset. In general, the MedMentions synthetic dataset has certain weaknesses due to its diversity.

5.4. Summary

In general, BC5CDR shows the best fidelity due to its simplicity, BioRED achieves a balance between diversity and structure, and MedMentions reveals the limitations of current LLMs when reproducing highly heterogeneous biomedical corpora. Using SDMS as a evaluation system, it can be deduced that greater diversity reduces the validity of the generated data, and large multi-domain corpora remain the most challenging scenario for the generation of synthetic biomedical texts. Therefore, temperature control in the generation of biomedical synthetic data is important but can be challenging due to complex datasets such as MedMentions.

6. Conclusions

This paper presents an initial analysis of synthetic biomedical sentence generation using knowledge-based constraint-guided beam-tree search, evaluated using the Synthetic Medical Data Scorecard (SMDS). Experiments conducted in three heterogeneous benchmark datasets (BC5CDR, BioRED, and MedMentions) show that SMDS provides a structured domain-oriented framework capable of revealing strengths and weaknesses in synthetic text generation that would not be captured by mere fidelity or diversity metrics.

This evaluation shows that low-temperature (0.01–0.2) produces more stable and reliable synthetic data (better consistency with real datasets or fewer structural errors). It also shows that the difficulty of synthetic generation increases with semantic diversity. BC5CDR, with only two domains, shows very good results in almost all criteria proposed by SMDS.

However, BioRED, which contains six domains and richer interactions between entities, maintains consistency and completeness, but shows an increase in constraint violations. MedMentions, the largest and most heterogeneous dataset with 21 domains, highlights the current limitations of synthetic data driven by LLM (reduced lexical coverage, increased structural inconsistencies, and difficulty reproducing the diversity of the original corpus, despite maintaining the complete presence of entities).

Finally, the evaluation presented here has certain limitations. The generated datasets have not been supervised by experts, all analyses are performed directly on the generated dataset without evaluating their biomedical accuracy. As a result, hallucinations or poorly generated sentences may not be noticed. The SMDS assesses structural, statistical, and distributional fidelity but does not evaluate biomedical correctness or clinical plausibility at the semantic level.

The results presented here are an initial demonstration of how SMDS can be applied in the generation of a synthetic biomedical dataset. Future work will incorporate additional generation strategies, human

intervention curation and evaluation with medical experts, and will also try with other LLMs. In addition, complementary evaluation frameworks that directly assess biomedical validity and reasoning consistency in synthetic biomedical datasets will be explored.

Acknowledgments

The publication of this article was funded by the University of Mannheim.

Declaration on Generative AI

We used generative AI to generate synthetic data. During the preparation of this work, the authors used Grammarly in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] A. Kumar, A. Gond, Natural language processing: Healthcare achieving benefits via nlp, ScienceOpen Preprints (2023).

[2] K. Jain, V. Prajapati, Nlp/deep learning techniques in healthcare for decision making, Primary Health Care: Open Access 11 (2021) 373–380.

[3] S. Sai, A. Gaur, R. Sai, V. Chamola, M. Guizani, J. J. Rodrigues, Generative ai for transformative healthcare: a comprehensive study of emerging models, applications, case studies, and limitations, IEEE Access 12 (2024) 31078–31106.

[4] M. Hofer, A. Kormilitzin, P. Goldberg, A. Nevado-Holgado, Few-shot learning for named entity recognition in medical text, arXiv preprint arXiv:1811.05468 (2018).

[5] M. Salmi, D. Atif, D. Oliva, A. Abraham, S. Ventura, Handling imbalanced medical datasets: review of a decade of research, Artificial intelligence review 57 (2024) 273.

[6] M. Delmas, M. Wysocka, A. Freitas, Relation extraction in underexplored biomedical domains: A diversity-optimized sampling and synthetic data generation approach, Computational Linguistics 50 (2024) 953–1000.

[7] L. Pilgram, S. El Kababji, D. Liu, K. El Emam, Magnitude and impact of hallucinations in tabular synthetic health data on prognostic machine learning models: Validation study, Journal of Medical Internet Research 27 (2025) e77893.

[8] K. Soman, P. W. Rose, J. H. Morris, R. E. Akbas, B. Smith, B. Peetoom, C. Villouta-Reyes, G. Cerono, Y. Shi, A. Rizk-Jackson, et al., Biomedical knowledge graph-optimized prompt generation for large language models, Bioinformatics 40 (2024) btae560.

[9] L. Long, R. Wang, R. Xiao, J. Zhao, X. Ding, G. Chen, H. Wang, On llms-driven synthetic data generation, curation, and evaluation: A survey, arXiv preprint arXiv:2406.15126 (2024).

[10] A. A. Barr, J. Quan, E. Guo, E. Sezgin, Large language models generating synthetic clinical datasets: a feasibility and comparative analysis with real-world perioperative data, Frontiers in Artificial Intelligence 8 (2025) 1533508.

[11] G. Grazhdanski, V. Vasilev, S. Vassileva, D. Taskov, I. Antova, I. Koychev, S. Boytcheva, Synthmedic: Utilizing large language models for synthetic discharge summary generation, correction and validation, Journal of Biomedical Informatics (2025) 104906.

[12] O. Litake, B. H. Park, J. L. Tully, R. A. Gabriel, Constructing synthetic datasets with generative artificial intelligence to train large language models to classify acute renal failure from clinical notes, Journal of the American Medical Informatics Association 31 (2024) 1404–1410.

[13] S. Shakeri, C. dos Santos, H. Zhu, P. Ng, F. Nan, Z. Wang, R. Nallapati, B. Xiang, End-to-end synthetic data generation for domain adaptation of question answering systems, in: Proceedings

of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 5445–5460.

[14] A. R. Mianroodi, A. Rezaie, N. G. Todorov, C. Rakovski, F. Rudzicz, Medsynth: Realistic, synthetic medical dialogue-note pairs, arXiv preprint arXiv:2508.01401 (2025).

[15] J. Wang, Z. Yao, Z. Yang, H. Zhou, R. Li, X. Wang, Y. Xu, H. Yu, Notechat: a dataset of synthetic patient-physician conversations conditioned on clinical notes, in: Findings of the Association for Computational Linguistics: ACL 2024, 2024, pp. 15183–15201.

[16] M. Nadăș, L. Dioșan, A. Tomescu, Synthetic data generation using large language models: Advances in text and code, IEEE Access (2025).

[17] Z. Li, H. Zhu, Z. Lu, M. Yin, Synthetic data generation with large language models for text classification: Potential and limitations, arXiv preprint arXiv:2310.07849 (2023).

[18] J. Lai, J. Zhang, S. Xu, T. Chen, Z. Wang, Y. Yang, J. Zhang, C. Cao, J. Xu, Llm-based automated theorem proving hinges on scalable synthetic data generation, arXiv preprint arXiv:2505.12031 (2025).

[19] R. Xin, C. Xi, J. Yang, F. Chen, H. Wu, X. Xiao, Y. Sun, S. Zheng, M. Ding, Bfs-prover: Scalable best-first tree search for llm-based automatic theorem proving, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025, pp. 32588–32599.

[20] G. Zamzmi, A. Subbaswamy, E. Sizikova, E. Margerrison, J. G. Delfino, A. Badano, Scorecard for synthetic medical data evaluation, Communications Engineering 4 (2025) 130.

[21] A. Kurakova, H. Homayouni, A comprehensive evaluation framework for synthetic medical tabular data generation, Journal of Biomedical Informatics (2025) 104939.

[22] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al., A survey on evaluation of large language models, ACM transactions on intelligent systems and technology 15 (2024) 1–45.