**Issue 17, 2025**

https://doi.org/10.23939/sisn2025.17.403

**УДК 004.8:811.161.2**

# SED-UA-SMALL: UKRAINIAN SYNTHETIC DATASET FOR TEXT EMBEDDING MODELS

**Oleksandr Mediakov[1], Dmytro Martjanov[2], Vasyl Lytvyn[3]**

[1]Lviv Polytechnic National University,
Department of Information Systems and Networks, Lviv, Ukraine
[2]Lviv Polytechnic National University,
Department of Artificial Intelligence, Lviv, Ukraine
[1]Email: oleksandr.mediakov.mnsam.2023@lpnu.ua, ORCID: 0000-0002-2580-3155
[2]Email: dmytro.i.martianov@lpnu.ua, ORCID: 0009-0003-3919-4412
[3]Email: vasyl.v.lytvyn@lpnu.ua, ORCID: 0000-0002-9676-0180



**This paper presents Small *S*ynthetic *E*mbedding *D*ataset, a fully synthetic dataset in Ukrainian designed for training, fine-tuning, and evaluating text embedding models. The use of large language models (LLMs) allows for controlling the diversity of generated data in aspects such as NLP tasks, asymmetry between queries and documents, the presence of instructions, support for various languages, and avoidance of social biases. A zero-shot generation approach was used to create a set of Ukrainian query-documents pairs with corresponding similarity scores. The dataset can be used to evaluate the quality of multilingual embedding models, as well as to train or fine-tune models to improve their effectiveness when working with Ukrainian texts. The paper covers a comprehensive description of the dataset construction process, including the parameters influencing the diversity of generated texts, the large language models used for actual generation of the data, and an example of using the dataset to evaluate and compare selected multilingual embedding models on the task of semantic text similarity. Unlike existing Ukrainian datasets, which are mainly based on real texts, SED-UA-small is fully synthetic, providing greater flexibility in controlling the diversity and specificity of data for the needs of training and evaluating embedding models, and allowing for feast and cost-effective expansion of the dataset with high-quality entries if needed. We used a combination of open and proprietary large language models of different sizes to generate the first version of the dataset, consisting of 112 thousand text pairs, divided into training (~50 %), testing (25 %), and validation (25 %) sets. The data is publicly available at https://huggingface.co/datasets/suntez13/sed-ua-small-sts-v1.**

**Keywords – text embedding, natural language processing, large language models, Ukrainian text processing.**

## Problem statement

Generating synthetic data using LLMs is a popular, efficient, and fast way to create high-quality datasets suitable for fine-tuning and improving text embedding models. The use of quality LLMs allows for controlling the diversity of generated data in terms of tasks (similarity, retrieval, QA, classification, and others), asymmetry (i.e., distinguishing queries from passages), adding instructions, supporting multiple languages, the presence or absence of social biases, and more. In this work, we propose an approach to generating a fully synthetic (zero-shot) dataset with a set of Ukrainian query-passage pairs – the Small **S**ynthetic **E**mbedding **D**ataset for the Ukrainian Language (SED-UA-small) – which can be used to evaluate the quality of multilingual embedding models, train or fine-tune models to improve their performance with

Ukrainian texts. The work includes a description of the main aspects of data construction, a description of the input parameters for diversifying the generated texts, access to the generated dataset published on the HuggingFace service, and an example of using this data to evaluate the quality and compare selected multilingual embedding models.

The availability of multilingual text embedding models that support the Ukrainian language might not be sufficient to ensure their versatility for solving NLP tasks with Ukrainian texts. This work proposes a method to address this problem by generating the first synthetic Ukrainian dataset for tuning embedding models that supports the asymmetry of queries and passages, instructions for instruction-following models, and has a certain diversity regarding NLP tasks, topics, and text formality.

## Analysis of Recent Studies and Publications

The development of high-quality text embedding models is a key direction in modern natural language processing, as they form the basis for a wide range of tasks, including semantic search, clustering, classification, and question answering. They are part of the RAG architecture and play an important role in creating memory mechanisms for LLM agents. Traditionally, the training of such models relied on large volumes of labeled or unlabeled real-world data (Muennighoff et al., 2023). However, with the advent of powerful large language models (LLMs), new opportunities have emerged for generating synthetic data that can be effectively used to improve the quality of existing and training new embedding models (Wang et al., 2024a).

In the current landscape of language models, there are two main approaches to generating synthetic data for embedding models: fully synthetic generation (where both queries and passages, as well as other dataset components, are generated) (Wang et al., 2024a), and few-shot generation (where some parts of the data, such as passages, come from real data, while others are generated by a language model) (Lee et al., 2024). One of the first successful cases of using the first type – fully synthetic generation – was demonstrated by the Wang et al. (2024a), they showed the effectiveness of using LLMs to generate pairs of queries and relevant documents that were then used to train embedding models, leading to a significant improvement in their performance on various benchmarks. Similarly, the Lee et al. (2024) showed that using a dataset distilled from an LLM can significantly improve the quality of embedding models with truncated embedding dims. This approach, and the corresponding dataset, is called the FRet (Few-shot Prompted Retrieval) dataset – which is a large-scale, synthetically generated dataset specifically designed to improve the retrieval capabilities of text embedding models. FRet is an example of the second type of synthetic dataset.

A separate application of datasets is the evaluation of the quality of embedding models, which is an important aspect of their development and application. Various benchmarks are used for this purpose, covering different NLP tasks such as relevant information retrieval, semantic text similarity, classification, and clustering, each supported by a corresponding dataset or several of them. The Massive Text Embedding Benchmark (MTEB) framework (Muennighoff et al., 2023) is one of the most comprehensive sets of benchmarks, allowing for a thorough evaluation of embedding models across many tasks. Its multilingual version, MMTEB (Enevoldsen et al., 2025), also includes the Ukrainian language, for which 8 datasets have been added for 4 types of tasks. At the same time, 303 datasets have been added for English, and a total of 228 tasks for European languages. Accordingly, it can be concluded that there is a current lack of benchmarking coverage for the Ukrainian language.

On the same time the work on Ukrainian-language datasets for many other NLP tasks is present. Most available resources are based on real texts obtained from websites, news articles, social networks, translated etc. (Chaplynskyi, 2023; Dementieva et al., 2025). While such datasets are valuable, they may be limited in terms of the diversity of represented tasks and the complexity of annotation (w.r.t. embedding models). Generating synthetic Ukrainian data using LLMs opens new possibilities for creating specialized datasets tailored to the specific needs of training and evaluating embedding models for the Ukrainian language, which is particularly relevant given the growing interest in Ukrainian-language NLP applications. The main contribution of our paper is the extension and expansion of the existed dataset to overall increase the performance of language models of different types over Ukrainian texts.

## Formulation of the Article's Objective

The main goal of this work is to develop and present a fully synthetic Ukrainian dataset (Synthetic Embedding Dataset for Ukrainian Language – SED-UA-small), specifically created for evaluating and improving the quality of multilingual text embedding models when working with the Ukrainian language. Unlike existing Ukrainian NLP resources, which are mainly based on real data, our approach involves using the power of large language models to generate diverse "query-passage" pairs, covering a wide range of semantic relationships and embedding usage scenarios.

Achieving this goal involves solving the following tasks: developing a methodology for generating synthetic Ukrainian data using LLMs, defining key parameters for controlling the diversity and quality of the generated data, creating and publishing the SED-UA-small dataset in open access, and demonstrating its practical application for evaluating the quality and comparing existing multilingual embedding models on Ukrainian texts. The results of this work should contribute to the development of Ukrainian NLP by providing researchers and developers with a potentially valuable resource for improving embedding models. Our main goal is to contribute the existed body-of-work related to the Ukrainian NLP resources and enhance it.

## Main Results

### Synthetic Embedding Dataset

To address the identified gap in Ukrainian-language datasets for evaluating and improving text embedding models, we developed an approach for synthetic data generation and used it to produce publicly available dataset. The core of our methodology lies in a looped usage of a one-step generation algorithm that leverages the capabilities of large language models. This process allows for the creation of diverse and task-oriented data by randomly sampling a set of generation parameters.

The one-step generation algorithm begins with the random formation of generation parameters for the dataset entries, which include: the NLP task, topic, number of positive texts, number of negative texts, and formality of style. The formality of style can take one of 5 values (formal, informal, more formal than informal, etc.). The number of positive and negative texts can vary from 1 to 10, and in the general case, these two numbers are not equal. For the topic 20 unique values were constructed, but then aggregated into 15 categories, including business and marketing, IT, medicine and health, cooking, culture, entertainment, art, etc. The list of supported NLP tasks that were proposed to the base LLM is as follows:

Information retrieving (IR).

Documents retrieving.

Question answering (QA).

Semantic similarity or sentence-to-sentence similarity (STS).

Paraphrase detection.

Natural Language Inference (NLI).

Accordingly, a randomly generated set of parameters was sampled and then used to format a prompt template, which was then input into the language model. The number of positive and negative passages were sampled from corresponding discrete uniform distributions, same holds for the task type. In order to increase the potential diversity of text w.r.t. topics the selection of one for a single generation process was constructed with the possibility to select more than 1 topic or none. With probability 0.2 instead of one - two topics were sampled (from uniform distribution), while in case of a single selection it was possible that none of the topics were sampled.

As a result of the generation, the model was required to generate the following entries:

Instruction for the query.

Query text.

Instruction for the passage (same for positive and negative examples).

*n* positive texts passages.

*k* negative text passages.

After this, the generated result underwent post-processing, including simple criteria of acceptance and the addition of metadata about the generative process. To form the final dataset format, the query and all passages were transformed into a set of text pairs with a corresponding similarity score (0 or 1).

The acceptance criterion for all generated entries was based on text diversity estimation using the Self-BLEU metric (Zhu et al., 2018). Self-BLEU, which equals the average BLEU score of all possible pairs of texts, was employed to control the diversity of the generated data. The underlying idea was to accept any generated entry where the Self-BLEU score, calculated between all pairs of each query and each generated passage, and each generated passage and every other generated passage, was lower than a p redefined threshold. A lower Self-BLEU score indicates higher text diversity because it signifies less n-gram overlap between the texts. In our paper that threshold was set to 0.9

Using this relatively simple approach of Self-BLEU allows the managing of the diversity of both generated negative and positive passages. It also helped prevent the issue where a generated positive passage could be an exact copy of the query. However, this method has inherent limitations. It cannot guarantee that the generated text is not degenerate or that the generated positive and negative passages are semantically similar or dissimilar to the query in the desired way. The addition of methods capable of controlling these crucial characteristics of the generated dataset entries will be explored in future research.

A visual representation of the simplified generation process of the SED-UA-small dataset is shown on Fig. 1.

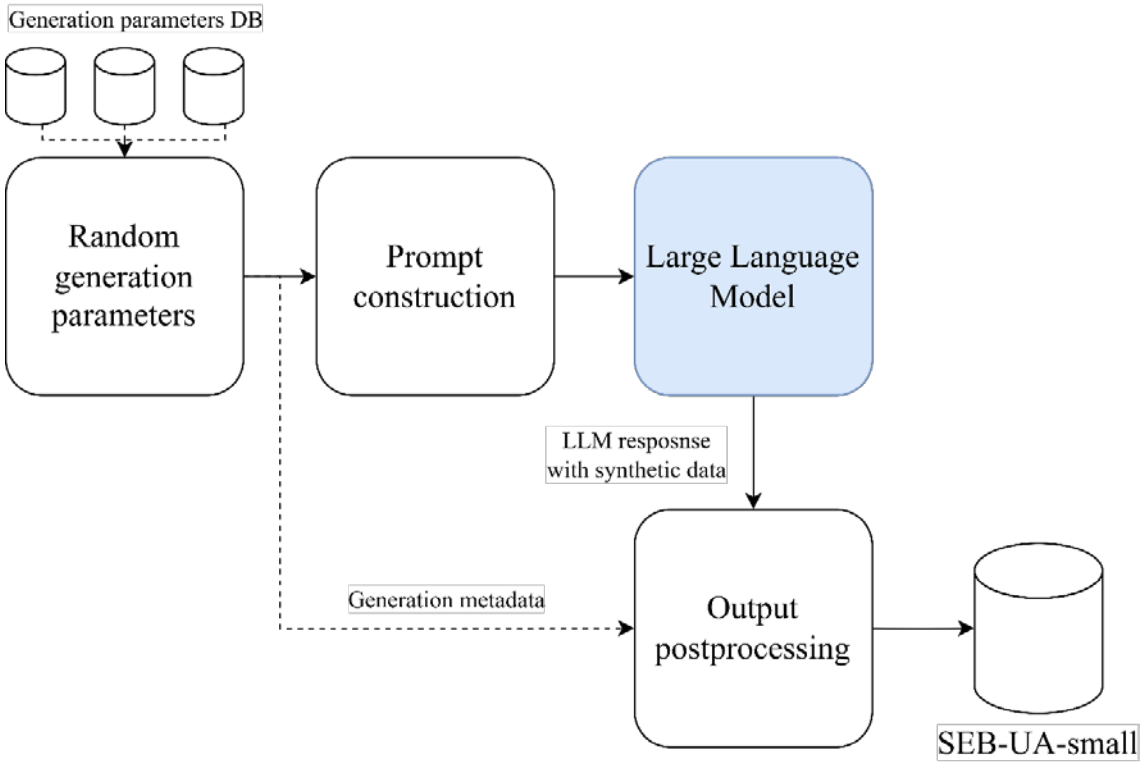


*Fig. 1. Simplified scheme for a generation step*

Obviously, the quality of the generated data directly depends on the base LLM that creates the texts. At the same time, to achieve speed and cost-effectiveness, it is necessary to use smaller (or cheaper) models. To balance price and quality during generation, open and proprietary models of various sizes were used.

The results from the following models were retained when creating this version of the dataset:

Gemini 2.0 (flash-001) (Pichai et al., 2024) – responsible for 73 % of the generated data.

Llama 3.3 (70B Instruct) (Grattafiori et al., 2024) – responsible for 20 % of the generated data.

Gemma 2 (27B) (Gemma Team et al., 2024) – responsible for the other 7 % of the data.

In total, SED-UA-small contains 112k rows of data, which were divided into three subsets – training (~50%), testing, and validation (25% each). Detailed statistics on the splits are shown in Table 1. During the splitting, it was ensured that different rows with the same query would only be in one of the subsets.

*Table 1*

**Statistics of the dataset splits**

| Split | # of rows | Scores distribution | | # of unique queries |
|---|---|---|---|---|
| | | # of positives | # of negatives | |
| train | 56713 | 27736 (49%) | 28977 (51%) | 2355 |
| test | 27942 | 13563 (49%) | 14379 (51%) | 1182 |
| validation | 27176 | 13282 (49%) | 13894 (51%) | 1145 |

The impact of the base LLM on text diversity can be estimated with distribution of the Self-BLEU for all accepted entries per model. Corresponding visualization can be seen on Fig. 2. As can be seen from Fig. 2 the diversity of the model by Google on average and by median has lower Self-BLEU per entry indicating diversity within generated queries and passages, while Llama model has lower diversity. It's important to note that we estimate the distribution of Self-BLEU for independent generated set from a single query and list of positive and negative passages, which means that Llama might not have diversity within single generated dataset entries, but it doesn't indicate low text quality overall.

As an example of using the created dataset, we propose the creation of a small benchmark for multilingual embedding models that support the Ukrainian language. For this purpose, a sentence similarity evaluation was conducted for the selected models, calculated as a Pearson correlation between expected and extracted cosine similarities. The full list of models used in the benchmark includes:

The *paraphrase-multilingual-MiniLM-L12-v2* and *distiluse-base-multilingual-cased-v1* models (Reimers & Gurevych, 2019).

The *universal-sentence-encoder* model (checkpoint name *use-cmlm-multilingual*) (Cer et al., 2018).

The LaBSE model from paper by Feng et al. (2022).

BGE M3-Embedding models (Chen et al., 2024).

Granite Embedding Models (107m and 278m versions) (Granite Embedding Team, 2024).

Multilingual E5 Text Embeddings (small, base and large versions) (Wang et al., 2024b).

The *test* split was used to calculate the correlation and for the resulting comparison. All models were used according to their specifications (i.e., necessary instructions were added to the texts if required). The result of the model's evaluation is shown in Fig. 2.

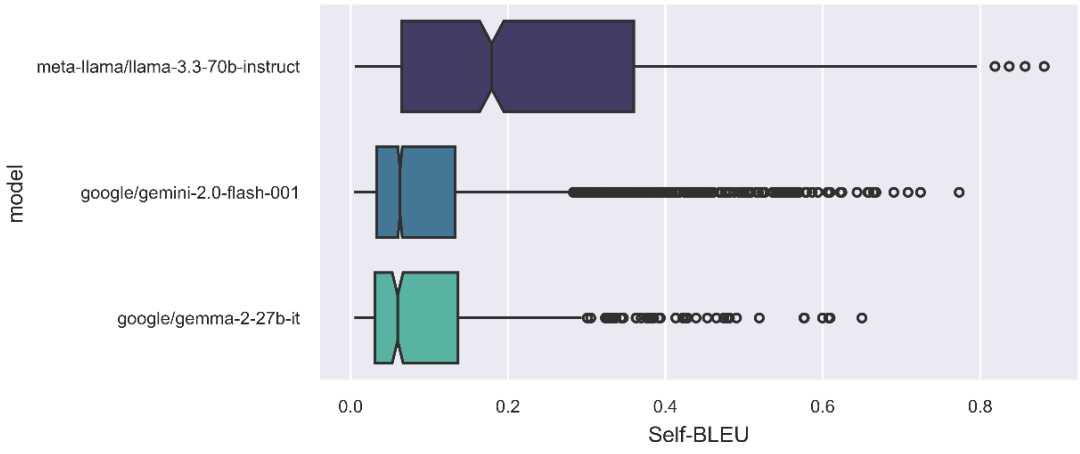


*Fig. 2. Distributions estimations of the Self-BLEU score per model*

As can be seen from Fig. 2, some of the models received a high correlation score, indicating their ability to work with Ukrainian text. A detailed analysis of the evaluation results is beyond the scope of this work. These initial results demonstrate the potential of SED-UA-small for evaluating multilingual embedding models on Ukrainian text.

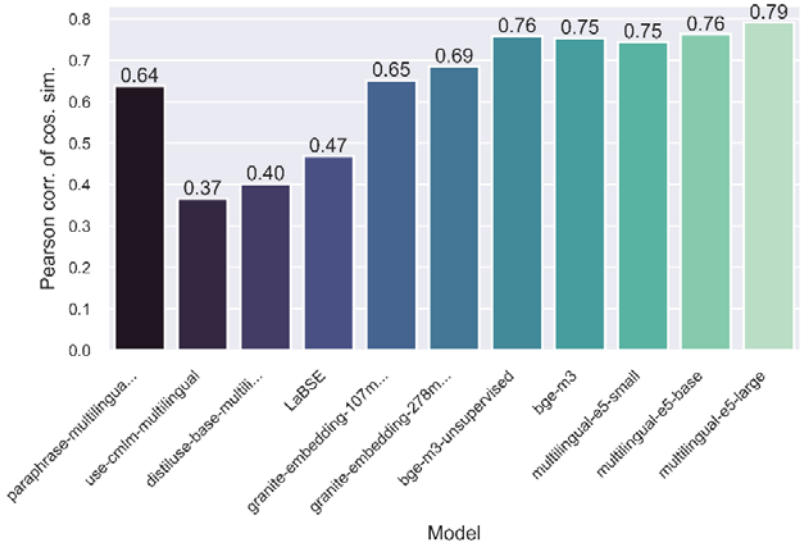


*Fig. 3. Evaluation results of embedding models on test split*

The current limitations of the generated dataset are primarily determined by the algorithm used for its creation and the base LLM. For example, the data contain examples of different positive and negative texts for a single query, but there are no passages that would have multiple examples of queries. Of course, it is possible to generate additional negative query pairs for a single passage or use existing techniques for that. Also, problems with hallucinations or the creation of degenerate text can impair the quality of some generated data. Therefore, it is necessary to also use real data for the full training or evaluation of embedding models.

The scientific significance of this work lies in its exploration of fully synthetic data generation as a viable and efficient methodology for creating added resources Ukrainian language. By demonstrating the ability of LLMs to produce a diverse dataset suitable for evaluating and potentially enhancing text embedding models, this research contributes to the broader understanding of how synthetic data can address data scarcity challenges in NLP. Furthermore, the introduction of SED-UA-small provides a novel benchmark for assessing the cross-lingual transfer capabilities of multilingual embedding models specifically concerning the Ukrainian language. This allows for a more nuanced understanding of model performance beyond aggregate multilingual benchmarks, potentially revealing language-specific strengths and weaknesses. The methodological details of the data generation process, including the controlled introduction of task diversity, asymmetry, and stylistic variations, offer valuable insights for future research in synthetic data creation for specialized NLP tasks and languages.

Our contribution in creating and openly publishing SED-UA-small on HuggingFace directly addresses the identified gap in Ukrainian NLP resources for instruction based embedding models, fostering further research and development in this domain by providing a tangible and accessible tool for the community. Our work expands and enhances the body-of-work related to the Ukrainian NLP resources.

## Conclusions

In this paper, we introduced SED-UA-small, the fully synthetic Ukrainian dataset designed for training and evaluating text embedding models. Our approach leverages the capabilities of LLMs to generate diverse query-passage pairs covering various NLP tasks, topics, and formality levels. The dataset supports the asymmetry between queries and passages and includes instructions for instruction-following models. We

detailed the data generation process, highlighting the parameters used to control diversity and the specific LLMs employed to balance quality and cost-effectiveness. The resulting dataset, comprising 112k text pairs, has been made publicly available on HuggingFace. Furthermore, we demonstrated the utility of SED-UA-small by conducting a preliminary benchmark of several multilingual embedding models, providing an initial assessment of their performance on Ukrainian semantic similarity tasks.

While our work represents a valuable resource for the Ukrainian NLP community, several avenues for future work exist. Further research could focus on expanding the dataset's size and diversity by incorporating a wider range of NLP tasks and topics, as well as exploring more sophisticated prompting strategies and leveraging larger or fine-tuned LLMs for data generation. Additionally, a more comprehensive evaluation of a broader set of multilingual and Ukrainian-specific embedding models is warranted, utilizing SED-UA-small in conjunction with existing real-world datasets to provide a more robust assessment. Investigating techniques to mitigate potential biases and artifacts introduced during the synthetic data generation process is also crucial. This work contributes to the growing body of research on synthetic data generation for NLP and provides a foundation for advancing the development and evaluation of text embedding models for the Ukrainian language.

## REFERENCES


1. Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., John, R. S., … Kurzweil, R. (2018). *Universal sentence encoder*. https://doi.org/10.48550/arXiv.1803.11175
2. Chaplynskyi, D. (2023). *Introducing UberText 2.0: A corpus of modern Ukrainian at scale*. 1–10. Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.unlp-1.1
3. Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., & Liu, Z. (2024). *BGE m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation*. https://doi.org/10.48550/arXiv.2402.03216
4. Dementieva, D., Khylenko, V., & Groh, G. (2025). *Cross-lingual text classification transfer: The case of ukrainian* (O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, & S. Schockaert, Eds.). Association for Computational Linguistics. Retrieved from https://aclanthology.org/2025.coling-main.97/
5. Enevoldsen, K., Chung, I., Imene Kerboua, Kardos, M., Mathur, A., Stap, D., … Ömer Çağatan. (2025). *MMTEB: Massive multilingual text embedding benchmark*. https://doi.org/10.48550/arXiv.2502.13595
6. Feng, F., Yang, Y., Cer, D., Naveen Arivazhagan, & Wang, W. (2022). *Language-agnostic BERT sentence embedding*. https://doi.org/10.48550/arXiv.2007.01852
7. Granite Embedding Team, IBM. (2024). Granite embedding models. Retrieved from https://github.com/ibm-granite/granite-embedding-models/
8. Grattafiori, A., Dubey, A., Abhinav Jauhri, Pandey, A., Abhishek Kadian, Al-Dahle, A., … Rao, A. (2024). *The llama 3 herd of models*. https://doi.org/10.48550/arXiv.2407.21783
9. Lee, J., Dai, Z., Ren, X., Chen, B., Cer, D., Cole, J. R., … Naim, I. (2024). *Gecko: Versatile text embeddings distilled from large language models*. https://doi.org/10.48550/arXiv.2403.20327
10. Niklas Muennighoff, Tazi, N., Magne, L., & Reimers, N. (2023). *MTEB: Massive text embedding benchmark*. https://doi.org/10.48550/arXiv.2210.07316
11. Reimers, N., & Gurevych, I. (2019, November). *Sentence-bert: Sentence embeddings using siamese BERT-Networks*. Association for Computational Linguistics. https://doi.org/10.48550/arXiv.1908.10084
12. Sundar Pichai, Hassabis, D., & Kavukcuoglu, K. (2024, December 11). Introducing Gemini 2.0: our new AI model for the agentic era. Retrieved from Google website: https://blog.google/technology/google-deepmind/google-gemini-ai-update-december-2024/
13. Team, G., Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Surya Bhupatiraju, … Tsitsulin, A. (2024). *Gemma 2: Improving open language models at a practical size*. https://doi.org/10.48550/arXiv.2408.00118
14. Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024a). *Improving text embeddings with large language models*. https://doi.org/10.48550/arXiv.2401.00368
15. Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024b). *Multilingual E5 text embeddings: A technical report*. https://doi.org/10.48550/arXiv.2402.05672
16. Zhu, Y., Lu, S., Zheng, L., Guo, J., Zhang, W., Wang, J., & Yu, Y. (2018). *Texygen: A benchmarking platform for text generation models*. Association for Computing Machinery. https://doi.org/10.1145/3209978.3210080

# SED-UA-SMALL: УКРАЇНОМОВНИЙ СИНТЕТИЧНИЙ НАБІР ДАНИХ ДЛЯ МОДЕЛЕЙ ВБУДОВУВАННЯ ТЕКСТУ

**Олександр Медяков[1], Дмитро Мартьянов[2], Василь Литвин[3]**

[1,3]Національний університет "Львівська політехніка",
кафедра інформаційних систем та мереж, Львів, Україна
[2]Національний університет "Львівська політехніка",
кафедра систем штучного інтелекту, Львів, Україна
[1]Email: oleksandr.mediakov.mnsam.2023@lpnu.ua, ORCID: 0000-0002-2580-3155
[2]Email: dmytro.i.martianov@lpnu.ua, ORCID: 0009-0003-3919-4412
[3]Email: vasyl.v.lytvyn@lpnu.ua, ORCID: 0000-0002-9676-0180



**У даній роботі представлено Small Synthetic Embedding Dataset, повністю синтетичний набір даних українською мовою, розроблений для навчання, донавчання та оцінки моделей вбудовування текстів. Використання великих мовних моделей дозволяє контролювати різноманітність згенерованих даних за такими аспектами, як NLP-задачі, асиметричність між запитами та документами, наявність інструкцій, підтримка різних мов та уникнення соціальних зміщень. При генерації набору даних було використано підхід без навчання на прикладах цільового завдання до генерації для створення набору пар запитів та відповідних їм текстів українською мовою. Набір даних може бути використаний для оцінки якості мультимовних моделей вбудовування текстів, а також для навчання або донавчання моделей з метою підвищення їхньої ефективності при роботі з україномовними текстами. Робота охоплює детальний опис процесу побудови набору даних, включаючи параметри, що впливають на різноманітність генерованих текстів, використані мовні моделі, а також приклад використання набору даних для оцінки та порівняння відібраних мультимовних моделей вбудовування текстів на задачі семантичної подібності текстів. На відміну від наявних україномовних наборів даних, які переважно базуються на реальних текстах, SED-UA-small є повністю синтетичним, що надає більшу гнучкість у контролі різноманітності та специфічності даних для потреб навчання та оцінки таких моделей, дозволяє швидко та економічно ефективно розширювати набір даних високоякісними записами. Ми використовували комбінацію відкритих та приватних великих мовних моделей різних розмірів для генерації першої версії набору даних, що складається з 112 тисяч пар текстів, розділених на тренувальний (~50 %), тестовий (25%) та валідаційний (25 %) набори. Дані доступні за посиланням - https://huggingface.co/datasets/suntez13/sed-ua-small-sts-v1.**

**Ключові слова – вбудовування текстів, опрацювання природньої мови, великі мовні моделі, опрацювання української мови.**