https://doi.org/10.31891/csit-2026-2-7

UDC 004.75

TOOLS FOR SEMANTIC SEARCH AND ANSWER GENERATION IN UKRAINIAN-LANGUAGE MUSEUM SYSTEMS

Khrystyna LIPIANINA-HONCHARENKO
Doctor of Technical Sciences, Associate professor
West Ukrainian National University
https://orcid.org/0000-0002-2441-6292
kh.lipianina@wunu.edu.ua
Vadym VITENKO
PhD student of the Department of Information and Computer Systems and Control
West Ukrainian National University
https://orcid.org/0009-0002-2824-1044
vadvit009@gmail.com
Diana ZAHORODNIA
PhD in Technical Sciences, Associate professor
West Ukrainian National University
https://orcid.org/0000-0002-9764-3672
dza@wunu.edu.ua

Received: 01/04/2026
Accepted:05/05/2026
Published:31/05/2026

This article examines modern semantic search models and a search-based response generation approach for building intelligent museum information systems focused on Ukrainian-language and bilingual textual materials. It is shown that traditional lexical search, based on keyword matching, does not always ensure adequate quality in cultural heritage tasks, where relevance is determined not only by formal term matching but also by the semantic proximity of the query and the document. In this context, text vector representation models take on particular significance, as they allow texts to be represented in vector space and enable the retrieval of contextually relevant information even in the absence of direct lexical matches. This paper analyzes the LaBSE, paraphrase-multilingual-MiniLM-L12-v2, multilingual-E5-large-instruct, and BGE-M3 models in terms of Ukrainian language support, multilingualism, search accuracy, computational resource requirements, and suitability for local deployment. The generation of search-based responses is examined separately as an approach in which a generative model constructs a response based on fragments retrieved by a search module from an external knowledge base. It is argued that this approach is particularly suitable for museum information systems, as it reduces the risk of hallucinations, improves the factual accuracy of responses, and enables a source verification mechanism. The advantages and limitations of open and closed generative models in the context of Ukrainian-language museum services are also systematized. Based on the analysis, practical recommendations are formulated regarding the selection of search models and the configuration of answer generation systems for digital cultural heritage. The practical significance of these findings lies in the fact that they can be used to design museum information kiosks, digital guides, visitor assistance systems, and services providing access to cultural heritage in the Ukrainian language. Further research should focus on an experimental comparison of search models using a specialized Ukrainian-language museum corpus, as well as on evaluating the reliability of sources and the accuracy of citations in real-world digital cultural heritage systems.

Keywords: semantic search, search-based answer generation, vector representations, multilingual models, Ukrainian language, museum information systems, digital cultural heritage.

Introduction

Semantic search is an approach to information retrieval that relies on comparing the meaning of texts rather than on exact keyword matches. Unlike traditional lexical search, such as BM25, where results depend on common words between the query and the document, semantic search uses vector representations of sentences and text fragments. Such representations make it possible to find contextually relevant information even in cases where there are no direct word matches between the query and the document, which is particularly important for languages with complex morphology, a wealth of word forms, and synonymic series, to which the Ukrainian language belongs.

Specialized models are used to build semantic search, these models convert text into multidimensional vectors, i.e., vector representations. The emergence of Sentence-BERT-class models had a significant impact on the development of this field, as they enabled the generation of high-quality semantic representations of sentences and substantially improved semantic search results [1]. Further development of this approach led to the emergence of multilingual models that support not only English but dozens or even hundreds of languages simultaneously. This makes it possible to build a unified vector knowledge base for documents in Ukrainian and English and to implement cross-

lingual search, where a query is formulated in Ukrainian but the relevant document may be in English, or vice versa [4, 5].

Another key concept in modern intelligent information systems is search-based answer generation. This approach combines semantic search with the capabilities of large language models to produce more accurate and meaningful responses based on external knowledge. First, the system retrieves relevant documents or fragments from the knowledge repository, and then the language model generates a response to the query based on the content of the retrieved documents. This approach reduces the risk of generating false statements and improves the accuracy of responses, as the model operates not only based on its own parametric knowledge but also utilizes up-to-date information from the knowledge base [3].

For Ukrainian museum projects that work with bilingual textual materials and require reliable, source-verified answers, semantic search and search-based answer generation systems are particularly important components. In such an environment, the system must not simply "know" general facts, but rely on museum descriptions, exhibit labels, archival records, historical materials, and other verified sources. This is precisely why a review of modern semantic search models with Ukrainian language support, as well as models and approaches for generating answers suitable for use in the museum domain, is relevant.

Analysis of recent research and problem formulation

The development of semantic search in recent years has significantly changed approaches to organizing information access in intelligent systems. While early search engines relied primarily on keyword matching and statistical relevance models, modern systems increasingly rely on dense vector representations of text. An important milestone in this development was the emergence of Sentence-BERT, which adapted the BERT architecture to the tasks of sentence comparison and semantic search [1].

Further research has led to the development of multilingual models capable of mapping texts in different languages into a shared semantic space. Of particular importance in this context is the LaBSE model, designed for multilingual alignment of sentences and phrases in over a hundred languages [4, 5]. Its use has opened up new possibilities for cross-lingual search, which is critically important for systems in which documents are stored in multiple languages.

Among the newer generation of search models, the E5 series and BGE-M3 models attract particular attention. The E5 series models are trained on a large corpus of text pairs using contrastive learning and fine-tuning on search tasks [6, 7]. They combine multilingualism, relatively high accuracy, and practical suitability for use in real-world systems. An even more versatile solution is the BGE-M3 model, which combines dense, sparse, and multi-vector search and supports long contexts [8, 9].

In parallel, the field of search-based answer generation has evolved. In a seminal work by P. Lewis et al., it was demonstrated that combining a search module with a generative model enables improved answer accuracy in knowledge-rich tasks [3]. Over time, the architectures of such systems have become more complex, but the basic principle has remained unchanged: the generative model should not operate in isolation from external information, but rather based on the relevant sources found.

Despite a significant body of research in the field of semantic search and search-based answer generation systems, the problem of selecting the optimal configuration for Ukrainian-language museum systems remains unresolved. It is necessary to consider not only the quality of search and generation but also factors such as local deployment, support for the Ukrainian language, working with a bilingual corpus, response latency constraints, and the ability to provide the user with source verification. This is why a systematic review of relevant models and the formulation of recommendations for their use in the museum domain are particularly relevant.

The aim of this article is to analyze modern models of semantic search and search-based response generation in terms of their suitability for building Ukrainian-language museum information systems, as well as to formulate practical recommendations regarding the selection of search and generative components for digital cultural heritage systems.

Presentation of the main material

Semantic search models are typically implemented as separate embedding models based on a transformer architecture. Such models independently encode queries and documents into vectors of fixed dimension. Next, using a similarity measure, such as cosine similarity, one can quickly find the most relevant documents for a given query. For practical implementation, vector representations are indexed in specialized vector databases, such as FAISS, Milvus, Qdrant, etc., which enables efficient nearest-neighbor search.

The first successful models of this type were predominantly English-language [1]. However, in recent years, powerful multilingual semantic search models have emerged that also cover the Ukrainian language. They are trained on multilingual corpora and parallel data in such a way that texts in different languages fall into a single shared semantic space. As a result, a query and a document in different languages can have similar vector representations if their content is equivalent [4, 5]. This creates the basis for cross-lingual search, where a user formulates a query in Ukrainian, and the system can find an English-language document containing a relevant answer.

The main current semantic search models that support the Ukrainian language are discussed below.

LaBSE (Language-Agnostic BERT Sentence Embedding) is a multilingual model developed by Google Research that can generate similar vectors for translated phrases in over 100 languages [4, 5]. The model is specifically trained on parallel texts to ensure the language independence of vector representations. It has become one of the powerful foundational solutions for cross-lingual search and semantic alignment across languages.

The advantage of LaBSE lies precisely in its multilingual versatility. It is particularly valuable for tasks where it is necessary to combine documents in many languages within a single semantic space, such as for translation search or cross-lingual search in large archives. In a museum context, this can be useful when some documents are in Ukrainian and others are in English, Polish, or other languages of international museum communication.

At the same time, LaBSE was introduced earlier than the latest search models, so it often falls short of more modern solutions in terms of accuracy. It remains important as a stable multilingual encoding model; however, in scenarios where the highest possible quality of semantic search is required, it is advisable to consider newer models. The model is open-source and available in the sentence-transformers/LaBSE repository [4]. It has approximately 500 million parameters, and the maximum input text length is limited to 256 sub-words.

Paraphrase-multilingual-MiniLM-L12-v2 is a lightweight multilingual model from the Sentence-Transformers family that supports over 50 languages, including Ukrainian [2]. It was obtained by fine-tuning a compact transformer architecture on multilingual paraphrases. The model generates 384-dimensional vector representations, and its parameter count is approximately 118 million.

The main advantage of this model is its high speed in generating vector representations and low memory requirements. This is why it is suitable for real-time systems, client applications, chatbots, prototypes, and other environments with limited computational resources. It can also be run on a central processing unit, which is practical for local information terminals and standalone museum installations.

A drawback is a certain reduction in accuracy compared to larger and more modern models. In particular, on complex, long, or multi-component queries, its ability to accurately distinguish relevant fragments is lower. However, in cases where speed is the top priority, paraphrase-multilingual-MiniLM-L12-v2 remains a perfectly reasonable choice.

Multilingual E5 is a series of open-source models from Microsoft, presented as an extension of the Sentence-T5 and E5 approaches [6]. The models in this series were trained on a large dataset comprising over a billion text pairs, using contrastive training followed by fine-tuning on retrieval tasks. The multilingual-e5-large-instruct variant covers 94 languages, including Ukrainian [7].

One of the key features of E5 is the use of a special input data format. Queries and documents are presented with prefixes such as query: and passage:, which helps the model better distinguish the role of the input text and improves the quality of the match between the query and the document. This approach is particularly useful for question-answering systems, knowledge base search, and other scenarios where a clear interpretation of the semantic role of the text is important.

Multilingual E5 is a balanced option for multilingual semantic search. It delivers high quality with moderate computational resource requirements and is well-suited for enterprise search, knowledge-based chatbots, museum information systems, and other applications where accuracy and practicality must be balanced. However, it should be noted that the model's context is limited to 512 tokens, so long documents must be pre-split into fragments [6, 7].

BGE-M3 (Multi-lingual, Multi-function, Multi-granularity Embedding) is one of the latest models for multilingual vector representations, introduced in 2024 [8, 9]. It supports over 100 languages, including Ukrainian, and demonstrates strong performance on many evaluation datasets for multilingual search.

The uniqueness of BGE-M3 lies in its versatility. The model is capable of generating not only dense vectors for semantic search, but also sparse representations that mimic lexical search, as well as multi-vector representations for handling long documents [8]. Architecturally, it is based on XLM-RoBERTa and supports a context length of up to 8,192 tokens. The model has approximately 540 million parameters, and the dimension of the vector representations is 1,024.

The practical benefit of using BGE-M3 is that it allows for the combination of semantic and exact keyword search within a single model. That is why the authors recommend using hybrid search as a combination of vector search and approaches such as BM25, as well as supplementing the processing sequence with a re-ranking stage to achieve maximum precision [8, 9]. For high-load production systems that work with multilingual corpora and require maximum comprehensiveness and accuracy, BGE-M3 is an extremely promising solution. A drawback of the model is its resource intensity. Generating 1024-dimensional vector representations, especially for long texts, requires more time and memory than more compact models. Therefore, its use is appropriate where quality is a priority rather than minimal response latency.

As shown in Table 1, modern models, particularly Multilingual E5 and BGE-M3, outperform previous solutions in multilingual search tasks. At the same time, the choice of a specific model depends on the use case. If CPU speed is critical, it is advisable to use MiniLM. If maximum search quality in a multilingual environment is required, BGE-M3 is a better choice. For museum applications, where the knowledge base is typically of moderate size, the Multilingual E5 model offers a reasonable compromise, providing high-quality search in Ukrainian and English with moderate hardware requirements.

To improve the accuracy of semantic search, a re-ranking stage is often used. In this processing sequence, a separate encoding model first identifies the top N candidates, and then a separate model—typically a joint encoding model—re-ranks these results by analyzing the "query–document" pair in greater depth. Unlike separate encoding models, which encode the query and document separately, joint encoding models process them together, allowing for better consideration of local relationships and detailed matches.

Table 1

Key characteristics of semantic search models for Ukrainian-language and multilingual museum systems

Model	Languages	Model Type	Size	Speed	License	Local Execution	Recommended Scenarios
LaBSE	100+ languages, including Ukrainian	separate encoding model based on BERT-large	~0.5 billion parameters, 768-dimensional vector	average	Apache 2.0	yes	cross-language search, multilingual text alignment, translation search
paraphrase-multilingual-MiniLM-L12-v2	50+ languages, including Ukrainian	separate encoding model	~0.1 billion parameters, 384-dimensional vector	high	Apache 2.0	yes	real-time search, chatbots, resource-constrained systems
multilingual-E5-large-instruct	94 languages, including Ukrainian	separate coding model, fine-tuned using instructions	~0.6 billion parameters, 1024-dimensional vector	average	MIT	yes	universal multilingual search, question-and-answer systems, corporate and museum knowledge bases
BGE-M3	100+ languages, including Ukrainian	a separate encoding model that combines dense, sparse, and multi-vector search	~0.55 billion parameters, 1024-dimensional vector, 8192-dimensional context	lower than that of compact models	MIT	yes	high-precision multilingual search, hybrid search, handling of long documents

In a multilingual environment, it is possible to use multilingual co-encoding models or specialized re-ranking models. For museum systems that generate search-based responses, this step is useful when it is necessary to filter out random matches and ensure that the top results are precisely those fragments that best match the user's query. This is particularly important for long, complex, or compound queries, where a basic search may return thematically similar but not the most accurate fragments.

At the same time, re-ranking is an additional computational operation, so it should be used selectively: for example, for complex factual queries, longer questions, discussion topics, or cases where the quality of the base search is insufficient.

Search-based answer generation is an approach in which a large language model generates an answer based on additional context obtained through search. In the classic implementation, a dense search module, the Wikipedia index, and a generative model were used for search, combined into a single processing sequence [3]. Subsequently, this concept evolved, and various modifications emerged, but the underlying idea remained unchanged: external knowledge is integrated into the process of constructing a response.

In the context of a museum, this means that the language model should not "guess" the answer based solely on its own parametric knowledge. Instead, upon receiving a query, it must first retrieve fragments of exhibit descriptions, archival records, biographical materials, or scholarly explanations via a search module, and then use these to formulate a response for the visitor. This ensures greater accuracy, reduces the risk of hallucinations, and makes it possible to cite the sources or fragments on which the response is based.

When building such systems for a Ukrainian-English multilingual environment, at least two basic conditions must be taken into account. First, the search model must be multilingual to find relevant materials regardless of the language of the query or document. Second, the generative model must be able to function correctly in Ukrainian, both in terms of understanding the query and constructing a natural response.

The most powerful generative models today are large language models trained on extremely large amounts of data. Most of them were not specifically trained on Ukrainian datasets, but thanks to multilingual web corpora, they have acquired a certain level of understanding of the Ukrainian language. In the museum domain, a generative model must not only possess language proficiency but also be capable of correctly interpreting the given context, generalizing it, and formulating a concise, accurate, and understandable response.

GPT-4 is one of the most powerful proprietary text-generation solutions, demonstrating a high level of performance in multiple languages, including Ukrainian [10]. The model possesses exceptional capabilities for

reasoning, generalization, and generating natural responses; therefore, when combined with a search module, it can deliver a high-quality user experience.

However, GPT-4 is a proprietary commercial system and is primarily accessible via an API. This implies dependence on external infrastructure, network connectivity, a paid access model, and limited control over the runtime environment itself. For museum kiosks or local standalone systems, this can be a significant limitation. Nevertheless, as an external generative service, GPT-4 can be useful during the prototyping phase or in scenarios where the highest possible quality is a priority.

In 2023, Meta introduced LLaMA-2—an open-source series of generative models with 7, 13, and 70 billion parameters [11]. Thanks to training on a large multilingual corpus, these models support the Ukrainian language to a certain extent. While they are inferior in quality to GPT-4, they have a key advantage: the ability to be deployed locally.

For museum assistant tasks, the 7B or 13B models are particularly interesting, as they can be deployed on a single server or workstation with a graphics processing unit (GPU) and, given high-quality context, generate sufficiently acceptable responses. In this case, the search module compensates for some of the limitations of the generative model itself by providing it with accurate factual material.

BLOOM is an open-source multilingual model with 176 billion parameters, trained on data from 46 languages [12]. Ukrainian was included in the training corpus, so the model can generate text in Ukrainian. BLOOM's strength lies precisely in its pronounced multilingualism and openness.

At the same time, BLOOM's enormous size significantly limits its practical use. Deploying such a model requires very significant computational resources, making it unsuitable for fast museum services, local information terminals, or systems with limited budgets. Therefore, it should be viewed more as a research or demonstration solution.

Another area of focus is specialized, smaller-scale multilingual generative models, notably the mGPT family from SberAI [13, 14]. These are GPT-like decoder models with 1.3 billion and 13 billion parameters, trained on 61 languages, including Ukrainian.

The advantage of mGPT is its relative ease of deployment, especially for the smaller versions. For example, the 1.3-billion-parameter model can run even on a central processing unit, making it attractive for embedded or standalone solutions. At the same time, the quality of generation from such models is significantly lower than that of large modern language models. Therefore, it is advisable to use them only when there is a very strong search component, where the model does not need to reason "on its own," but only concisely and accurately convey the found context.

Table 2 shows that there is a clear trade-off between closed and open models. Closed models, particularly GPT-4, provide the highest generation quality but require an external programming interface. Open models offer more control over the infrastructure and can run locally, but they typically yield lower overall quality. For a museum project, it makes sense to start with a medium-sized open model, such as LLaMA-2 13B, and, if necessary, transition to a hybrid scenario in which a more powerful external model is used for critical queries.

A summary of the characteristics of semantic search and response generation models presented above shows that the choice of a specific configuration for a museum system is determined by a trade-off between performance, multilingual support, search accuracy, the ability to handle long documents, and the possibility of local deployment. To systematize these dependencies, Figure 1 presents a diagram for selecting semantic search and response generation models based on the requirements of a Ukrainian-language digital museum heritage system.

Table 2

Comparative characteristics of generative models in the context of museum generation systems supplemented by search

Model	Languages	Type	Size	Speed	License	Local	Use Cases
GPT-4	multilingual, including Ukrainian	decoder-based transformer	very large, closed model	accessible via a software API	closed	no	highest-quality generation, complex dialogues, accurate responses when external access is available
LLaMA-2	multilingual support	decoder-based transformer	7B / 13B / 70B	medium	open-ended with constraints	yes	local search-based response generation systems, dialogues, and question-answering systems of medium complexity
BLOOM	46 languages, including Ukrainian	decoder-based transformer	176B	low	open with restrictions	limited	research-oriented multilingual systems; not ideal for fast services
mGPT	61 languages, including Ukrainian	GPT-like decoder model	1.3B / 13B	high for smaller models	open-source	yes	small standalone bots, embedded solutions, resource-constrained systems

The diagram (Figure 1) demonstrates that for systems with limited resources, it is advisable to use paraphrase-multilingual-MiniLM-L12-v2; for cross-lingual search—LaBSE; for maximum accuracy and handling large text fragments—BGE-M3, while the most balanced solution for the museum domain is multilingual-E5-large-instruct.

Regarding response generation, the figure illustrates the choice between closed, high-precision models, such as GPT-4, and open models suitable for local execution, specifically LLaMA-2 and mGPT. In conclusion, the recommended configuration for museum applications involves combining an effective search component, a re-ranking stage, and a generative model capable of producing concise and factually sound results.

Based on this review, a number of practical recommendations can be formulated for building a semantic search system with answer generation in the museum sector.

First and foremost, it is advisable to use multilingual vector representations for the search component. The most balanced option is a model such as the multilingual E5 or a similar one that supports Ukrainian and English. It ensures a high recall rate among relevant materials regardless of the language of the query or document. For a small corpus of museum texts, a medium-sized model can operate locally with sufficient speed and accuracy. If the volume of data increases or quality requirements become more stringent, it is advisable to switch to a more powerful model, such as BGE-M3, and implement a hybrid search that combines semantic search and keywords.

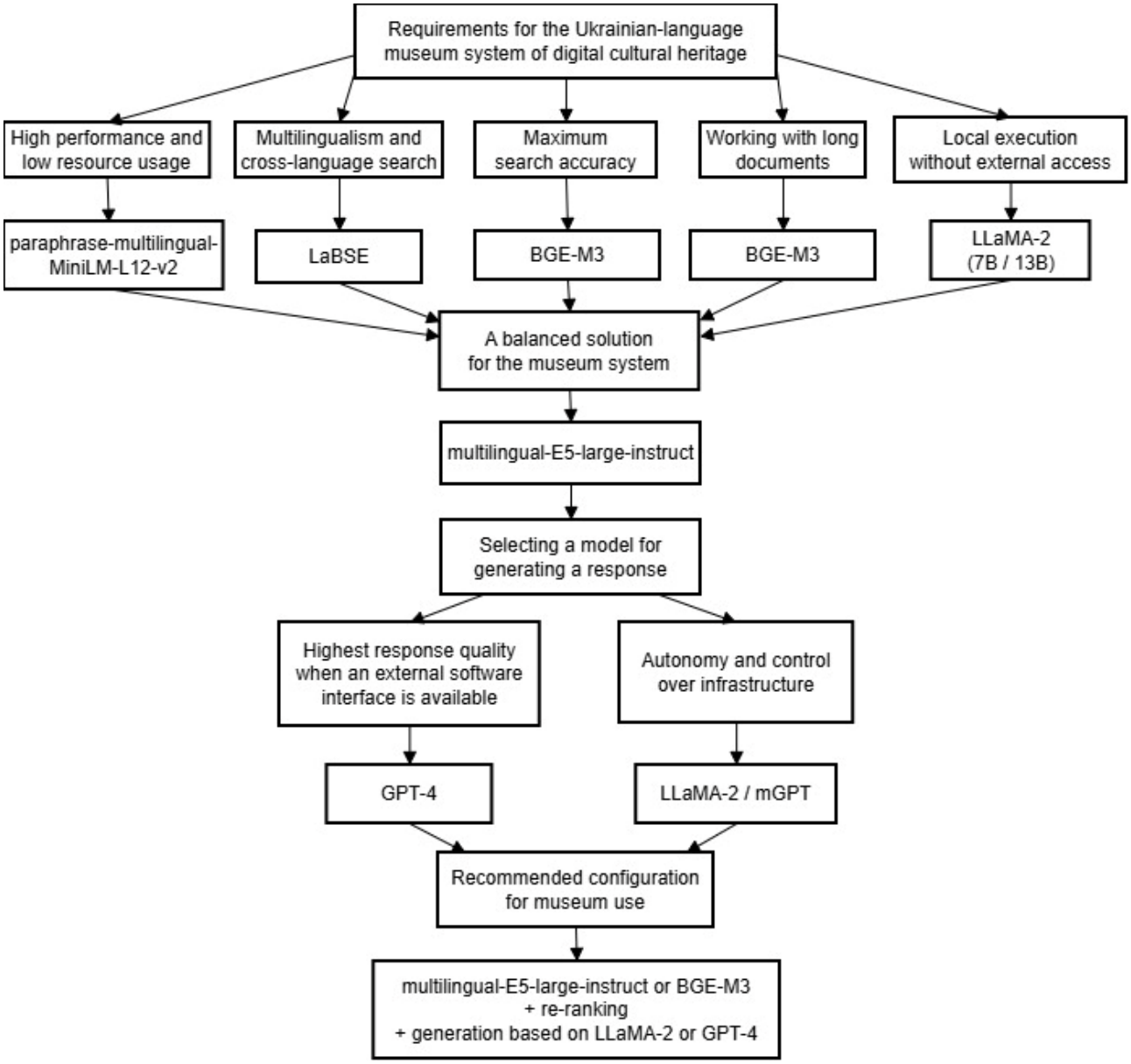

Fig. 1. Diagram illustrating the selection of semantic search models and the generation of responses based on the requirements of a Ukrainian-language digital museum heritage system

The second important aspect is knowledge structuring and information filtering. Before building a vector index, it is advisable to store document metadata: date, source, document type, exhibition, exhibit, language, etc. This will allow the search to be limited to a relevant subset of data. For a museum, this means that the system can search for results only within a specific collection, exhibition, period, or language segment, rather than across the entire knowledge base. This approach improves search accuracy and reduces the risk of accidental or irrelevant results.

The third important step is to add re-ranking for complex queries. If the search sometimes returns thematically similar but not the most accurate results, integrating a multilingual re-ranking module will improve the order of results. It is advisable to use re-ranking not for all queries, but only for highly complex cases, so as not to unnecessarily increase latency.

Another key requirement is limiting the length of the response and supporting citations. Museum visitors typically expect a concise answer of 1–3 sentences, rather than a lengthy narrative. Additionally, it is important to provide the ability to link to a source, such as an exhibit label, a scholarly reference, or an archival text. This increases trust in the system and allows users to verify the factual basis of the response.

Separately, the issue of local execution should be considered. Semantic search can already be performed locally today, even on modest hardware. Generation is the more complex part, but the 7B–13B models can be run on a workstation with a graphics processing unit. If such hardware is not available, the system should be designed so that the generative component is modular: during the prototype phase, it can operate via a software interface, and later be replaced with a local model without having to rebuild the entire architecture.

Finally, any museum system for generating search-based responses must be evaluated using real-world data. After implementation, it is advisable to measure Recall@5, Recall@10, MRR, or nDCG on a collected set of typical user queries. It is also necessary to verify the quality of the final answers in terms of factual support from the found documents, i.e., to assess the validity of sources, the correctness of citations, and the proportion of statements that lack confirmation in the found context.

Conclusions

This article provides a systematic review of semantic search models and search-based response generation approaches in the context of developing Ukrainian-language museum information systems. It is shown that traditional keyword search does not provide sufficient quality in cases where relevance is determined by semantic proximity rather than formal lexical similarity. In this regard, multilingual vector representation models should form the basis of modern museum search systems.

The main modern semantic search models with Ukrainian language support are analyzed. It is established that LaBSE remains useful for cross-lingual alignment, paraphrase-multilingual-MiniLM-L12-v2 is suitable for fast and lightweight systems, multilingual E5 demonstrates an optimal balance between accuracy and performance, and BGE-M3 is the most promising option for high-precision multilingual and hybrid search scenarios.

We also consider a search-based response generation approach as a foundation for building intelligent museum assistants and digital guides. It is shown that in the museum domain, the key factor in quality is not so much the power of the generative model itself, but rather the effectiveness of the search component, which must provide the generative component with accurate and relevant context. It is the search that should serve as the core of trust in the system, while generation acts as a natural language interface to the retrieved knowledge.

The practical significance of these findings lies in the fact that they can be used to design museum information kiosks, digital guides, visitor assistance systems, and services providing access to cultural heritage in the Ukrainian language. Further research should focus on an experimental comparison of search models using a specialized Ukrainian-language museum corpus, as well as on evaluating the reliability of sources and the accuracy of citations in real-world digital cultural heritage systems.

ADDITIONAL INFORMATION

AUTHOR CONTRIBUTIONS

Conceptualization, K. L.-H.; methodology, V.V.; validation, D.Z.; formal analysis, V.V.; investigation, V.V.; data curation, D.Z..; writing-original draft preparation, V.V.; writing-review and editing, K. L.-H.; visualization, V.V.; supervision, K.L.-H.; project administration, D.Z. All authors have read and agreed to the published version of the manuscript.

DECLARATION ON THE USE OF GENERATIVE ARTIFICIAL INTELLIGENCE TOOLS

In preparing this work, the author used DeepL Translate and Grammarly for: grammar and spelling checks, paraphrasing, and rephrasing. After using these tools/services, the author reviewed and edited the content and takes full responsibility for the content of this publication.

[1] Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 3982–3992.

[2] Sentence-Transformers. (2026). *paraphrase-multilingual-MiniLM-L12-v2*. Hugging Face. URL: https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

[3] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP

Tasks. *Advances in Neural Information Processing Systems*, *33*, 9459–9474.

[4] Sentence-Transformers. (2026). *LaBSE*. Hugging Face. URL: https://huggingface.co/sentence-transformers/LaBSE

[5] Feng, F., Yang, Y., Cer, D., Arivazhagan, N., & Wang, W. (2022). Language-agnostic BERT Sentence Embedding. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 878–891.

[6] Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R., & Wei, F. (2024). Multilingual E5 Text Embeddings: A Technical Report. *arXiv preprint*, arXiv:2402.05672.

[7] intfloat. (2026). *multilingual-e5-large-instruct*. Hugging Face. URL: https://huggingface.co/intfloat/multilingual-e5-large-instruct

[8] 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. *arXiv preprint*, arXiv:2402.03216.

[9] BAAI. (2026). *bge-m3*. Hugging Face. URL: https://huggingface.co/BAAI/bge-m3

[10] OpenAI. (2023). GPT-4 Technical Report. *arXiv preprint*, arXiv:2303.08774.

[11] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., et al. (2023). Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint*, arXiv:2307.09288.

[12] InfoQ. (2022). *BigScience Releases 176B Parameter AI Language Model BLOOM*. URL: https://www.infoq.com/news/2022/07/bigscience-bloom-nlp-ai/

[13] ai-forever. (2026). *mGPT*. GitHub. URL: https://github.com/ai-forever/mgpt

[14] ai-forever. (2026). *mGPT*. Hugging Face. URL: https://huggingface.co/ai-forever/mGPT

Христина ЛІП'ЯНІНА-ГОНЧАРЕНКО, Вадим ВІТЕНКО, Діана ЗАГОРОДНЯ
Західноукраїнський національний університет

ІНСТРУМЕНТИ ДЛЯ СЕМАНТИЧНОГО ПОШУКУ ТА ГЕНЕРАЦІЇ ВІДПОВІДЕЙ В УКРАЇНОМОВНИХ МУЗЕЙНИХ СИСТЕМАХ

У цій статті розглядаються сучасні моделі семантичного пошуку та підхід до генерації відповідей на основі пошуку для побудови інтелектуальних музейних інформаційних систем, орієнтованих на україномовні та двомовні текстові матеріали. Показано, що традиційний лексичний пошук, заснований на зіставленні ключових слів, не завжди забезпечує належну якість у завданнях культурної спадщини, де релевантність визначається не лише формальним зіставленням термінів, але й семантичною близькістю запиту та документа. У цьому контексті моделі векторного представлення тексту набувають особливого значення, оскільки вони дозволяють представляти тексти у векторному просторі та дають змогу отримувати контекстуально релевантну інформацію навіть за відсутності прямих лексичних збігів. У цій статті аналізуються моделі LaBSE, paraphrase-multilingual-MiniLM-L12-v2, multilingual-E5-large-instruct та BGE-M3 з точки зору підтримки української мови, багатомовності, точності пошуку, вимог до обчислювальних ресурсів та придатності для локального розгортання. Окремо розглядається генерація відповідей на основі пошуку як підхід, у якому генеративна модель конструює відповідь на основі фрагментів, отриманих пошуковим модулем із зовнішньої бази знань. Стверджується, що цей підхід особливо підходить для музейних інформаційних систем, оскільки він знижує ризик галюцинацій, покращує фактичну точність відповідей та дозволяє реалізувати механізм перевірки джерел. Також систематизовано переваги та обмеження відкритих та закритих генеративних моделей у контексті україномовних музейних сервісів. На основі аналізу сформульовано практичні рекомендації щодо вибору моделей пошуку та конфігурації систем генерації відповідей для цифрової культурної спадщини. Практичне значення цих висновків полягає в тому, що їх можна використовувати для проектування музейних інформаційних кіосків, цифрових путівників, систем допомоги відвідувачам та сервісів, що забезпечують доступ до культурної спадщини українською мовою. Подальші дослідження мають бути зосереджені на експериментальному порівнянні моделей пошуку з використанням спеціалізованого україномовного музейного корпусу, а також на оцінці надійності джерел та точності цитувань у реальних системах цифрової культурної спадщини..

Ключові слова: семантичний пошук, генерація відповідей на основі пошуку, векторні представлення, багатомовні моделі, українська мова, музейні інформаційні системи, цифрова культурна спадщина.