UDC 004.58 + 004.94
DOI: 10.15587/1729-4061.2026.351414

DEVISING AN APPROACH TO CONSTRUCTING A SPECIALIZED DICTIONARY TO TRAIN CHATBOTS WITH GENERATIVE ARTIFICIAL INTELLIGENCE

Olha Kryazhych
Corresponding author
PhD, Senior Researcher, Associate Professor*
E-mail: economconsult@gmail.com
ORCID: https://orcid.org/0000-0003-1845-5014
Viacheslav Riznyk
Doctor of Pedagogical Sciences, Associate Professor, Professor**
ORCID: https://orcid.org/0000-0002-6083-2242
Vasyl Vasenko
PhD, Associate Professor, Head of Department
Department of Theory and Techniques of Technology Education and Computer Graphics***
ORCID: https://orcid.org/0000-0002-2527-6359
Vasyl Yakuba
PhD, Associate Professor**
ORCID: https://orcid.org/0000-0002-2228-8522
Kateryna Iushchenko
Doctor of Philosophy (PhD), Junior Researcher*
ORCID: https://orcid.org/0000-0001-5183-816X
Oleksii Kuprin
Doctor of Philosophy (PhD), Junior Researcher*
ORCID: https://orcid.org/0000-0002-3730-4759
Oleksandr Tsyrul
PhD Student*
ORCID: https://orcid.org/0009-0002-5945-5918
*Institute of Telecommunications and Global Information Space of the National Academy of Sciences of Ukraine
Chokolivskyi blvd., 13, Kyiv, Ukraine, 03186
Department of Digital Methods of Teaching*
***Hryhorii Skovoroda University in Pereiaslav
Sukhomlynsky str., 30, Pereiaslav, Ukraine, 08401

This paper investigates the process that builds a subject-specific training dictionary for training a chatbot involving generative artificial intelligence. The task addressed is to reproduce the structured vocabulary characteristic of the relevant subject area from subject-specific knowledge when interacting with a chatbot.

The result of this study is the construction of a model for the process that sequentially manages independent user requests. The model made it possible to estimate the mathematical expectation of the stage number at which the processing of the request by the chatbot is completed.

Based on the constructed mathematical model, linear and logical-probabilistic models for building a specialized dictionary have been proposed. The linear model searches for a combination of words by sequentially searching for terms. The result of this approach is the comparison of a keyword from the query with the corresponding term or word form from the dictionary. The logical-probabilistic model is based on the target cell – a probable word from the user's query. This is explained by the possibility of defining a word that agrees with the term of the XML dictionary and has maximum relevance to the user query.

A methodology and algorithm for building a specialized dictionary have been suggested. The tests made it possible to obtain average signature values for the response at an error of 0.004%, as well as ensure the stability of the results. In practice, this could be used under the conditions of forming a probability distribution of possible word combinations for generating a response.

The proposed approach could be used in practical tasks of chatbots' domain adaptation, in particular at project support portals and in scientific libraries, as well as for improving intelligent dialog systems focused on the formation of refined user queries

Keywords: large language model, subject-specific knowledge, terminology management, semantic consistency

Received 10.11.2025
Received in revised form 14.01.2026
Accepted date 23.01.2026
Published date 27.02.2026

How to Cite: *Kryazhych, O., Riznyk, V., Vasenko, V., Yakuba, V., Iushchenko, K., Kuprin, O., Tsyrul, O. (2026). Devising an approach to constructing a specialized dictionary to train chatbots with generative artificial intelligence. Eastern-European Journal of Enterprise Technologies, 1 (2 (139)), 58–67. https://doi.org/10.15587/1729-4061.2026.351414*

1. Introduction

The widespread introduction of generative artificial intelligence and language models in the form of chatbots has led to their increased use in technical, engineering, and information systems. Such systems are used to support decision-making, automate data analysis, and provide intelligent support to complex processes. At the same time, universal language models demonstrate limited effectiveness in highly specialized subject areas [1]. This is due to incorrect interpretation of terminology and insufficient contextual consistency of responses [2]. In addition, systems based on generative artificial intelligence and language models increasingly perform the functions of information retrieval systems [3]. This is

due to a rethinking of approaches to the selection and search for information, although systems with generative artificial intelligence cannot be attributed to classical information retrieval systems [4].

Systems with generative artificial intelligence do not simply search for information. They perform semantic interpretation and generate responses in natural language form. At the same time, universal language models used in chatbots do not always provide sufficient accuracy when searching and reproducing specialized information, especially in technical and science-intensive subject areas [5]. To improve the quality of functioning of artificial intelligence as an information search system, it is necessary to use specialized dictionaries that play the role of formalized sources of domain-oriented knowledge. Such dictionaries make it possible to specify the semantics of queries, reduce the ambiguity of terms, and ensure semantic consistency of search results and response generation in the process of training language models [6, 7]. From this point of view, three main factors are essential for specifying chatbots based on artificial intelligence as an information search system. The first is the definition of the types of information resources within which specialized information is searched. The second is responsible for the method of forming and interpreting queries using domain-oriented terminology. The third factor is the form of representation of results generated by the chatbot based on specialized dictionaries and language models.

One of the promising directions for increasing the efficiency of generative chatbots is the use of specialized dictionaries as structured sources of domain-oriented terminology. For example, based on ontologies [8, 9]. Such dictionaries make it possible to formalize the conceptual apparatus of the subject area, reduce the variability of language interpretations, and ensure the stability of the language model's responses during training and operation.

Analysis of modern approaches to training chatbots proves that the construction of specialized dictionaries for the needs of generative artificial intelligence is mostly carried out without taking into account the specificity of knowledge representation in language models. In this regard, an urgent scientific and technical task is to devise an approach to building specialized dictionaries to train chatbots with generative artificial intelligence, focused on increasing the accuracy and subject-matter adequacy of language models.

2. Literature review and problem statement

In [10], the construction of ontology frameworks for describing a complex subject area is considered. It is noted that large language models (LLMs), such as ChatGPT (USA), DeepSeek (China), and KIMI (China), together with curated databases such as Rice-Alterome and PubAnnotation, offer new opportunities for semi-automated curation of ontologies. However, this process is based on human review of sources and therefore is laborious and difficult to scale. This is due to the need to combine or refine description features. This problem can be solved by prompt-based queries for LLMs. However, this requires the construction of specialized vocabularies for training LLMs.

In [11], the research is aimed at processing and searching industrial LLM data through natural language (NLP) queries. The innovation is based on the construction of a model that allows chatbots with generative artificial intelligence to find answers to specialized terminology and context. The specified problem is solved by special tuning of the model with LLM based on augmented search generation (RAG). This approach requires intervention in LLMs. However, the work does not consider the built model through terminology management approaches and semantic consistency in queries, which excludes its rapid adaptation to other areas of knowledge.

In [12], the issue of specialization of chatbots is solved using intelligent agents. The proposed mechanism improves search using chatbots by effectively managing semantic complexity and routing queries between specialized knowledge bases. However, the issue of constructing datasets for training LLMs for specific user needs remains unresolved.

In work [13], LLMs are trained on data samples. The proposal helps solve the problem of specialization of search directions. However, the model built by the authors is quite complex. And the formation of samples requires access to resources with limited access.

In [14], the accuracy of ChatGPT 3.5 responses to questions related to medical information about disease symptoms was analyzed. Standard and non-standard terminology was used. Responses to questions with non-standard NLP terminology were less accurate. Based on this, conclusions were drawn about the inexpediency of using ChatGPT 3.5 to clarify symptoms. The issue of specialized training of the chatbot remained unresolved.

In [15], the issue of the number of words in the user's query and the use of stable terminology to obtain an accurate answer was resolved. The researchers conducted a number of experiments with ChatGPT 4. A conclusion was drawn about the increase in model performance. However, the question remained unresolved – how to improve the perception of natural human language by a chatbot.

Work [16] is entirely aimed at building the structure of queries and responses when using a chatbot in document development. The basis of training is a knowledge base. The issue of training a chatbot to recognize and perceive specialized terms from a knowledge base remains unresolved.

In [17], the problem of training LLMs on structured data is addressed. The research was conducted on small data sets. The results showed a small number of errors in the generated documents. The issue of forming specialized data sets for fine-tuning LLMs remains unresolved.

In [18], a basis for constructing a subject-specific lexicon for designing complex systems in energy based on LLM is proposed. This is achieved by multi-level term extraction and synonym expansion to ensure the practicality of the lexicon. A dictionary model with hint templates is used. The disadvantage of the study is the use of multi-level templates, which, due to their complexity, move the query construction away from NLP.

In [19], the relationship between the appearance of words in a sentence is used to train LLM. The basis of the innovation is a dictionary, the feature of which is the similarity of human habits of using natural language. The result is formed according to the semantic vector of connections. The disadvantage of the study is the complex structure of the dictionary, which is due to the peculiarities of the Chinese language.

Based on the above, the main question of our study is to find a solution to the construction of specialized dictionaries for training chatbots with generative artificial intelligence. This will make it possible, through training, to expand the capabilities of LLMs and use chatbots in working on projects

on the management of complex systems and applications in research in the basic sciences. Solving the task could make it possible to improve the algorithms for formalizing queries posed in natural language. In addition, the solution to construct specialized dictionaries for training chatbots would make it possible to expand the functions of chatbots and improve the productivity of LLMs.

3. The aim and objectives of the study

The purpose of our study is to devise an approach to constructing a specialized dictionary to train chatbots with generative artificial intelligence. The basis of this approach is the formalization of data sets (definitions, terms) from the field of subject-specific knowledge. These sets are further formalized in the form of a subject-specific training vocabulary. Such formalization should make it possible, with step-by-step prompting, to obtain the most accurate and meaningful answer to the user's query.

The purpose of the work will be accomplished through the following tasks:

– to build a model for the process of sequential independent user queries;

– to propose a linear model for the construction of a specialized dictionary;

– to propose a logical-probabilistic model for the construction of a specialized dictionary;

– to define a methodology for the construction of a specialized dictionary to train chatbots with step-by-step prompting.

4. The study materials and methods

The object of our work is the process that constructs a subject-specific training dictionary to train a chatbot with generative artificial intelligence. This scheme should reflect the lexical ordering of words from the domain of subject-specific knowledge typical for the subject area of the query. A query is a linguistic construct formulated in the user's natural language. The query has a certain stylistic organization, logical connections between the elements of the statement, and the presence of emotionally colored or neutral language statements. To this end, prompts are provided in the subject-specific training dictionary.

The principal hypothesis assumes that the effectiveness of LLM training based on a subject-specific training dictionary depends on the depth of processing of the user query. In other words, how accurately the NLP terms used by the user are identified by the chatbot. To do this, it is necessary to ensure the lexical ordering of words in the dictionary of subject-specific knowledge.

The work is simplified by using ChatGPT 5.2 Plus (Instant) for testing. Separate studies on the constructed version of the dictionary were conducted with other chatbots, but no substantive studies were performed.

Our work assumes that the chatbot user communicates with generative artificial intelligence to search for information from subject-specific knowledge. The user considers the chatbot as a higher-level information search system.

Based on this assumption, the construction of a subject-specific training dictionary for training chatbots with generative artificial intelligence plays an important role in the tasks of pattern matching. In this study, pattern matching is considered not in the narrow sense of classical string search, but as a generalized mechanism for matching user queries, domain texts, and terminological units with predefined linguistic and semantic templates. Such templates are built on the basis of a specialized dictionary and reflect established ways of using terms in a certain subject area.

A specialized dictionary in this context serves as a controlled set of reference language samples used to identify, normalize, and interpret domain-oriented vocabulary. The use of pattern matching makes it possible to match natural language queries by users with the corresponding terminological structures of the dictionary, taking into account the variability of formulations, stylistic differences, as well as the possible presence of emotional coloring. This ensures correct domain interpretation of queries even before the response generation stage by the language model [18, 20].

Using pattern matching mechanisms in the process of forming and applying specialized dictionaries makes it possible to increase the accuracy of semantic matching between incoming queries and the knowledge embedded in the model. As a result, the risk of incorrect generation of responses is reduced, the level of so-called hallucinations is reduced, and the use of domain terminology is ensured more stable. Thus, pattern matching is an important methodological component of the approach to training and adapting chatbots with generative artificial intelligence in specialized subject areas [21].

The task of pattern search in this work assumes the assignment of a certain pattern (specialized term, keyword) and finding the corresponding word in the subject-specific learner's dictionary (SSLD). SSLD consists of possible keywords and permissible standard words of subject-specific knowledge.

The mechanism of the leading prompt is based on two properties of SSLD:

– lexicographic ordering of words;

– information redundancy of SSLD.

The following notations are accepted:

– $A_j = (a_1,...a_i,...a_n)$ is the j-th word in SSLD; $i=\overline{1,n}$, $j=\overline{1,N}$;

– j – ordinal number of the SSLD word;

– n – word length;

– q – volume of the set of characters (alphabet) from which words and the query text are constructed.

The lexicographic ordering of words in a subject-specific learner's dictionary is defined as follows: each value of the symbol a_i is assigned a number k_i such that $1 \le k_i \le q$. In other words, each symbol a_i is assigned the ordinal number of this symbol in the alphabet of the set of symbols q. A subject-specific learner's dictionary is considered lexicographically ordered if the following condition is met

$$\sum_{i=1}^{n}\left(\alpha_{k_i}\cdot q^{n-i}\right)_j \rangle \sum_{i=1}^{n}\left(\alpha_{k_i}\cdot q^{n-i}\right)_{j+1}. \quad (1)$$

In practice, this means that words interpreted as numbers in the positional number system with base q are ordered in descending order of their values. It also follows from (1) that $k_1 > k_2 > ...$, i.e., it is assumed that the symbols of the alphabet are renumbered in decreasing order of their seniority.

The information redundancy of words in a subject-specific learner's dictionary means that out of all possible q^n combinations of n symbols, only N combinations are used to represent the actually existing words in the dictionary, which constitute a small part of q^n. This property makes it possible

to identify the desired reference word by part of the symbols of the sample word. In particular, this can be done by its initial part (the determinant of the reference word). In addition, on this basis, it is possible to implement an advanced prompt to the user in the form of a selected subset of words containing a given sample. In this case, verification of the sample set and identification of the reference are carried out visually.

To conduct experimental testing of this approach, the user must sequentially enter the symbols of the sample a_1, a_2, ..., (starting from the first). When entering a query or refining it, words with the same values of the corresponding symbols a_1, a_2, ..., a_k are selected from the subject-specific learner's dictionary to the current (virtual) reference book.

The set of words with the same values of the symbols a_1, a_2, ..., a_k is a set of size N_k. In this case, $N_{k+1} \le N_k$, i.e., the search area for the sample narrows as the symbols a_k are entered. The lexicographic ordering of the words of the subject-specific learner's dictionary accelerates the construction of the k-set. In the ideal case, with a strictly regular and uniform structure of the dictionary, the relation $N_k = \frac{N}{q^k}$ holds.

In practice, the distribution of N actually existing values of standard words among q^n of all possible combinations of symbols is random, and, accordingly, the N_k values are also random.

It is assumed that the words of the $(k + 1)$-set are a subset of the k-set. In addition, a sample word belonging to the $(k + 1)$-set also belongs to the k-set. The user's query can be constructed with respect to the use of standard words as linear or logical-probabilistic. An example of a linear query with refinements is described in [22]. In the context of using specialized dictionaries, a logical-probabilistic query to a chatbot is considered as a structured set of terms and logical connections between them. For such a query, the generative model calculates the probabilistic relevance of the answers taking into account domain-oriented terminology.

Accordingly, a specialized dictionary can be built using a linear or logical-probabilistic model. In the event of the described queries, the construction of the dictionary using a step-by-step prompt strategy can be built by specifying the given query. When entering the next symbol (word) a_k from the current directory with a volume of N_k words, the chatbot refers to the analysis and identification of a sample of directory terms selected according to a certain criterion, with a volume of m words. If the desired sample is not in this sample, and the user enters the next symbol a_{k+1}, then the search area is selected taking into account the entered word N_{k+1}. The chatbot performs the identification of the entered query until the desired sample is detected or until the fact of its absence in the dictionary is established (if the sample is not in the k-set of power N_k). In the latter case, certain actions can be performed that depend on the purpose of the chatbot training system with generative artificial intelligence. For example, this may be the continuation of the input of sample symbols and the replenishment of the subject-specific learner's dictionary.

To enable machine processing and integration with generative artificial intelligence systems, a specialized dictionary is represented in XML (eXtensible Markup Language) format. This format allows for hierarchical structuring of terms, their definitions, domain attributes, and semantic relationships between concepts, which is critically important for automated analysis and search. The XML-record of the dictionary in relation to this work provides a clear separation of data from the processing logic and supports semantic consistency. This is due to the fact that in the context of chatbots with generative artificial intelligence, XML-dictionaries are used as structured external sources of knowledge. They can be automatically converted into internal representations (tokens, embeddings, or contextual fragments) and involved at the stage of response construction.

Our study was performed using general-purpose personal computing systems. The hardware configuration included an Intel Core i7 central processor (USA), 32 GB of DDR4 RAM, an NVIDIA GeForce RTX 3060 graphics processor with 12 GB of video memory (USA), and a 1 TB NVMe SSD solid-state drive. These resources ensured the execution of tasks to build a specialized dictionary and conduct experiments with language models.

The Windows 11 Pro operating system (Microsoft, USA) software was used. The systematization of test results was carried out in MS Excel. The main programming language was Python 3.10 (USA). The NumPy 1.24, Pandas 2.0, and spaCy 3.7 libraries (USA, Germany) were used in the study. To experimentally test the effectiveness of the specialized dictionary, large language models were used via the OpenAI API (USA). The construction and validation of dictionary resources were carried out using the Visual Studio Code 1.85 (USA) environment.

5. Results of research on devising an approach to constructing specialized dictionaries to train chatbots

5. 1. Model of the process of sequential independent user queries

When constructing a query from subject-specific knowledge to a chatbot, the query is refined under certain conditions. This process can be considered as a sequence of tests, each of which can end with a successful or unsuccessful result. If the test under a condition i is unsuccessful, the corresponding condition is replaced by the next condition $i + 1$, after which the test is repeated. In the case of a successful result, the process is terminated.

The probability of a successful test result under condition i is equal to p_i and does not depend on the probabilities of the test results under the previous conditions $i - 1$, $i - 2$,... .

The tests are repeated many times with the same fixed sequence of conditions. If within the next series all n tests are unsuccessful, the process is terminated regardless of the subsequent conditions.

In the described case, it is necessary to determine the mathematical expectation for the value corresponding to the moment of completion of the process. That is, to determine the moment when the chatbot will complete the work with SSLD on processing the request.

To solve this problem, unconditional probabilities q_i of successful completion of the process at the stage corresponding to condition i are introduced. Since the trials are independent:

$$P_i^{\delta y} = P_i \prod_{s=1}^{i-1} \left(1 - P_{i_s}\right),$$

for $i = 1,..., n-1$,

$$P_n^{\delta y} = P_n \prod_{s=1}^{n-1} \left(1 - P_{i_s}\right) + \prod_{s=1}^{n} \left(1 - P_{i_s}\right).$$

As a result, the sum of the unconditional probabilities of all trial outcomes leading to the completion of the query identification process is equal to 1. The process ends under the condition $S_1 \vee S_2 \vee ... \vee S_n \vee \overline{S}_n = 1$, where S with the index is the user's clarification regarding the query topic.

Accordingly, it should be proved that

$$P^{\delta y}[n]= \\ =\sum_{i=1}^{n}P_i\prod_{s=1}^{i-1}\left(1-P_{i_s}\right)+ \\ +\prod_{s=1}^{n}\left(1-P_{i_s}\right)=1. \quad (2)$$

We assume $P^{\delta y}[n] = 1$, in this case, $P^{\delta y}[n+1] = 1$

$$P^{\delta y}[n+1]=\sum_{i=1}^{n+1}P_i\prod_{s=1}^{i-1}\left(1-P_{i_s}\right)+\prod_{s=1}^{n+1}\left(1-P_{i_s}\right)= \\ =P^{\delta y}[n]=1.$$

The validity of equation (2) for the length of the word $n = 3$ shows

$$P^{\delta y}[3]=P_1+P_2\left(1-P_1\right)+P_3\left(1-P_1\right)\left(1-P_2\right)+ \\ +\left(1-P_1\right)\left(1-P_2\right)\left(1-P_3\right)=1.$$

That is, the expression for determining the average value of i, which corresponds to the last step of completing the request processing process, takes the following form

$$\bar{i}=\sum_{i=1}^{n}iP_i\prod_{s=1}^{i-1}\left(1-P_{i_s}\right)+n\prod_{s=1}^{n}\left(1-P_{i_s}\right). \quad (3)$$

Model (3) demonstrates the mathematical expectation of the stage number at which the chatbot completes processing of a user request. It takes into account the probability of successful completion at each step given the failure of all previous stages, as well as the limiting case of forced completion after passing the entire sequence of checks.

5. 2. Linear model of a specialized dictionary

A linear model of SSLD can be considered as some abstract register that holds N cells of unit length. The cells are located along a straight line and numbered according to their coordinates. Each cell corresponds to a word of length n, formed from an alphabet of volume q, where q determines the system of representation of dictionary words, and n is the number of characters in a word.

Part of the cells, the number of which is denoted by M, is active. For active cells, the corresponding combinations of characters correspond to the actually existing words of the base dictionary. The placement of active cells among all cells of the register is random. For parameters N, q, and n, the condition is satisfied under which the space of possible combinations significantly exceeds the number of active words.

From the above assumptions, it follows that the probability r that a randomly selected cell of the dictionary with parameters N, q, and n is active is equal to the ratio of the number of active cells to the total number of possible cells, i.e.

$$r=\frac{M}{N}, \quad (4)$$

and the value of r is small.

A graphical interpretation of the proposed model is shown in Fig. 1.

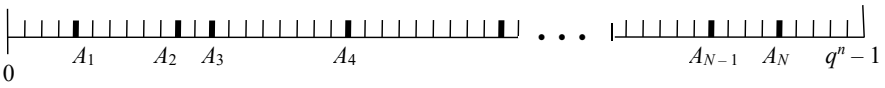


Fig. 1. Linear model of word distribution in a subject-specific learner's dictionary

Within the framework of our study, the active cells of the probabilistic model correspond to the terms recorded in a specialized XML dictionary, while relation (4) characterizes the density of domain-oriented terminology in the space of possible lexical combinations.

5. 3. Logical-probabilistic model of a specialized dictionary

In model (4), one of the active cells is chosen as the target cell TA. This cell corresponds to the searched word of the specialized dictionary under the condition of specifying a certain pattern. Each TA is assigned the value p_i, and $\Sigma\ p_i = 1$. This value is interpreted as the probability that when the dictionary is randomly accessed, cell i is the searched one. With lexicographic ordering of the dictionary, the p_i values are distributed along the register randomly (Fig. 2).

The arrangement of active cells in descending p_i values (Fig. 3) occurs by conditionally moving them to positions s.

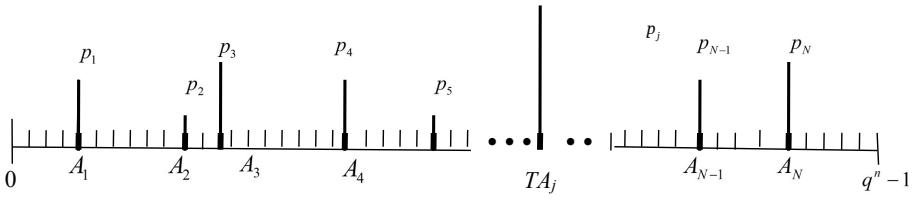


Fig. 2. Hypothetical probability distribution of queries to a subject-specific learner's dictionary

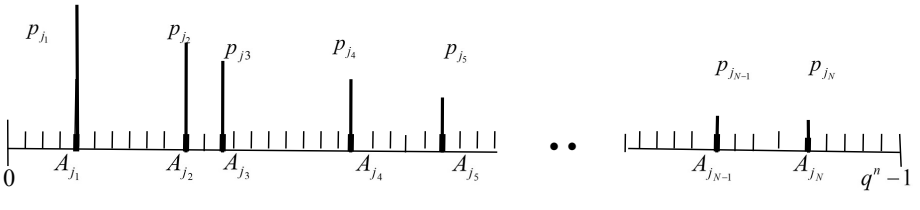


Fig. 3. Ordered probability distribution of queries

The position numbers (Fig. 3) correspond to numbers of the p_i values in the list, sorted in descending order. The resulting discrete probability distribution of p_i is fitted to a continuous function $p(x)$ (Fig. 4).

After fitting, a point x_0 is selected on the x-axis. The function $p(x)$ satisfies the condition $\int p(x)\,dx = 1$. On the x-axis, q points with coordinates x_φ ($\varphi = 1, 2,..., q^n$) are also selected (Fig. 4). The number π_φ is set in correspondence with each point:

$$\pi_\phi=\int_{x_{\phi-1}}^{x_\phi}p(x)\,dx; \\ \sum_{\phi=1}^{q^n}\pi_\phi=\int_0^L p(x)\,dx=1, \quad (5)$$

where the quantity L determines the length of the event space (or search).

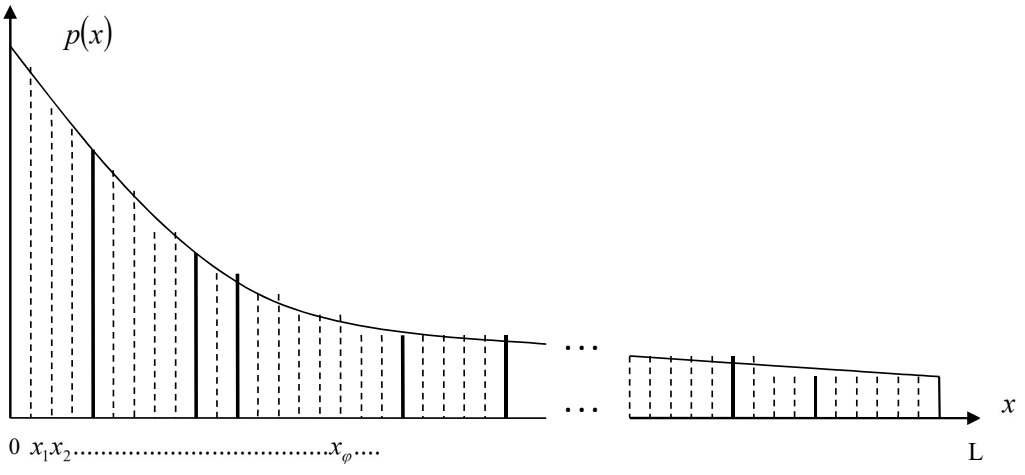


Fig. 4. Fitting an ordered probability distribution

The quantity r in model (5) is interpreted as the probability that for an arbitrary SSLD with parameters N, q, n and with an arbitrary query, cell i is an active and at the same time a searched cell.

When searching using a logical-probabilistic model, the process of finding a term in SSLD can be considered probabilistic. That is, a refinement word is entered step by step and an answer is obtained taking into account this word. At the beginning of the input process, there are q^n words, each of which is assigned the value π_φ. From the entered words of the first query, a word that corresponds to the searched active cell is randomly selected with a probability of π_φ.

At the next step i (i = 1, 2,...) from q^{n-i+1} words, $q^{n-i} - 1$ words are added randomly and equally likely to refine the query. The random word remains, and $q^{n-i} - 1$ words are entered for searching in the SSLD. Accordingly, the set g_i is used for searching. With a random word, the system with generative artificial intelligence processes $m_i = g_i + 1$ words. Then, m_i words are sorted by decreasing π_φ values. The system with generative artificial intelligence forms a response from m words with the largest π_φ values. If this response contains a random word selected earlier, then the process is considered complete. Otherwise, $i := i + 1$. That is, the described step is repeated until the random word is identified, and an answer is obtained with it.

quantitative measure of the effectiveness of using a specialized XML dictionary in the RAG architecture.

In the case of a high density of relevant terms (increasing p_i values for the top positions), the average value of i decreases, which corresponds to an earlier successful completion of the query processing. On the contrary, with a low density of specialized terminology, the process requires a larger number of iterations and may be forced to end after the permissible number of steps is exhausted. In this case, the construction of a specialized dictionary to train chatbots can be implemented according to the scheme in Fig. 5.

The construction of SSLD (Fig. 5) occurs according to the following algorithm:

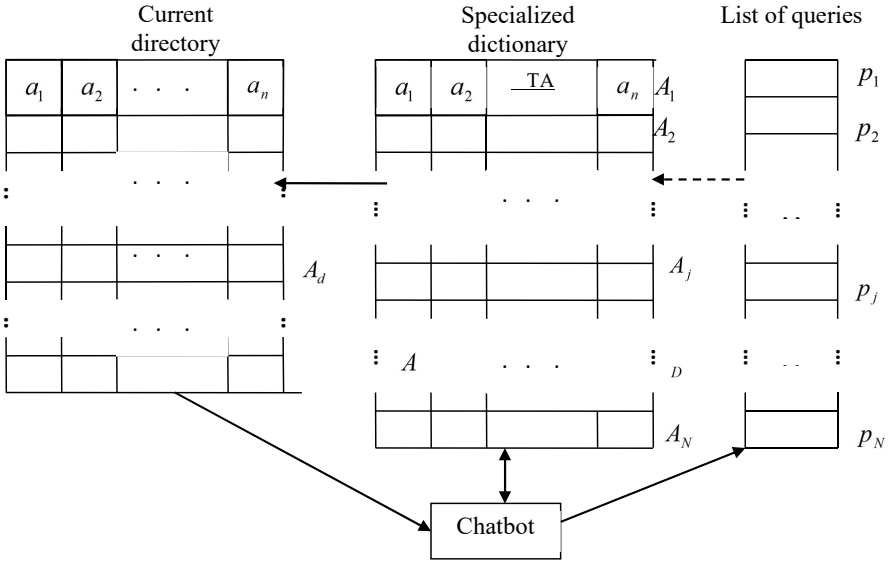


Fig. 5. Scheme for constructing a subject-specific learner's dictionary

5. 4. Construction of a dictionary with a prompt by steps

Taking into account (5), the probability p_i of successful completion of the process at step i is interpreted as the probability that the searched TA occurs in the answer from the words that have the largest p_i values and are presented to the language model for generating the answer. Accordingly, the product of model (5) characterizes the probability that at all previous stages of the search the searched term was not included in the formed context.

Thus, the mathematical expectation of the step number of the process completion according to model (3) determines the average depth of processing of the user query and serves as a

Step 1. Fixation and analysis of the initial query.

The user query is normalized (ordered). Key terms are highlighted. The initial vector of the semantic query is formed. The goal is to obtain an initial idea of the subject area.

Step 2. Construction of the current directory.

A subset of terms semantically close to the query is selected from the XML dictionary. The terms are ranked in descending order of p_i. As a result, the current directory is obtained according to an ordered list of words describing the essence of the query.

Step 3. Control over the list of clarification requests.

The history of the user's clarification requests is analyzed. Repeated, contradictory, or clarifying formulations are identified. The p_i values are adjusted taking into account the accumulated context. User clarifications are considered as additional observations in the probabilistic model.

Step 4. Generate a step-by-step prompt.

If the probability of successful completion of p_i is insufficient, a prompt is generated. The prompt is formulated as a short, guided query that specifies the terminology, object, or process, as well as level of detail. The prompt is added to the dialog as the next step.

Step 5. Dictionary update.

Terms with high p_i values are fixed as active dictionary elements. New or refined terms are added to the XML dictionary with the corresponding attributes. Less relevant elements are lowered in ranking or excluded from the current reference. As a result, the dictionary evolves in accordance with real usage scenarios.

Step 6. Checking the completion condition.

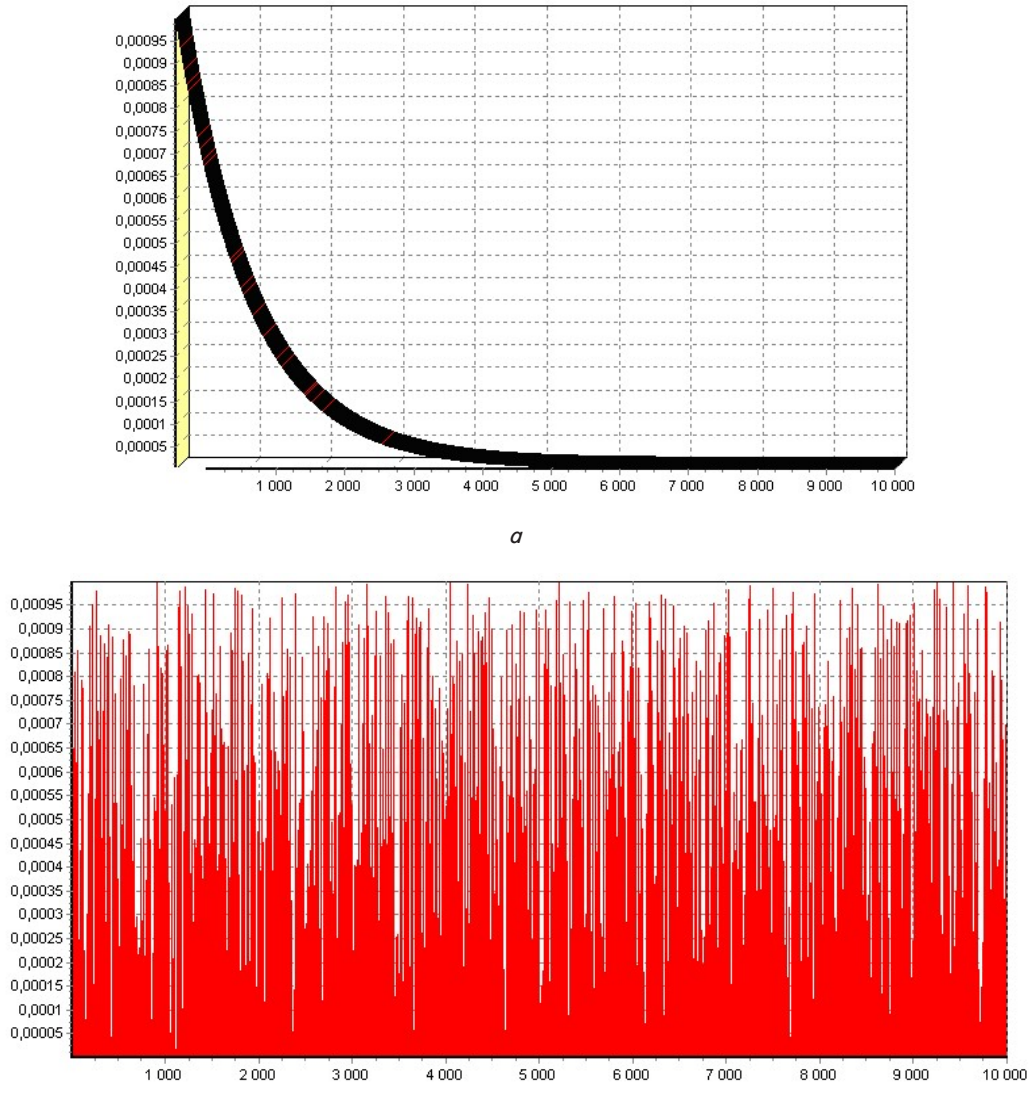
If the TA is included in the current portion of terms, the process is completed. The average number of the completion step i is calculated. The final chatbot response is formed.

According to the above scheme (Fig. 5), a SSLD was built and the probability distribution of queries with a step-by-step prompt for training the chatbot was tested. The specialized area of the SSLD is nuclear physics.

8 basic thematic variants of queries with the following parameters were investigated:

1) $q = 32$, $n = 8$, $N = 1.1 \cdot 10^6$, where $r = 10^{-6}$;
2) $q = 32$, $n = 8$, $N = 1.1 \cdot 10^4$, where $r = 10^{-8}$;
3) $q = 32$, $n = 8$, $N = 1.1 \cdot 10^3$, where $r = 10^{-9}$;
4) $q = 32$, $n = 8$, $N = 1.1 \cdot 10^2$, where $r = 10^{-10}$;
5) $q = 10$, $n = 12$, $N = 1 \cdot 10^6$, where $r = 10^{-6}$;
6) $q = 10$, $n = 12$, $N = 1 \cdot 10^4$, where $r = 10^{-8}$;
7) $q = 10$, $n = 12$, $N = 1 \cdot 10^3$, where $r = 10^{-9}$;
8) $q = 10$, $n = 12$, $N = 1 \cdot 10^2$, where $r = 10^{-10}$.

The results $p_j^{(0)}$ and p for the exponential distribution with parameters $N = 10000$, $r = 10^{-8}$ are shown in Fig. 6.



a

b

Fig. 6. Distribution of queries with step-by-step instructions for training a chatbot: *a* – probabilities $p_j^{(0)}$; *b* – probabilities p_j after hashing

Thus, hashing exponentially distributed requests acts as a mechanism for aligning the probability space, which is appropriate when building step-by-step chatbot training systems and reducing the information entropy of input data.

6. Discussion of the results of devising an approach to constructing a specialized dictionary

The model of completing a chatbot's query to a specialized dictionary (3) actually shows the course of one complete stage of attaining subject-specific knowledge. The model describes the mathematical expectation of the step number i, at which processing a user request by a chatbot with generative artificial intelligence ends in the presence of sequentially applied conditions or checks. In the proposed model, each step i corresponds to the next stage of processing a user request, in particular:

– checking the compliance of the request with the terms of the specialized dictionary. This gives an advantage over work [10], in which the processing of terms is performed by a person;

– clarifying the semantic context. This positively distinguishes our result from papers [11, 12], in which the semantic content remains a priori devalued;

– applying domain-oriented restrictions or attempting to generate a relevant response. This provides significant advantages in improving the operation of chatbots, in contrast to studies [18, 19], due to the possibility of constructing a system of coordinated prompts.

The probability p_i characterizes the successful completion of the request processing at the i-th step, that is, the situation in which a correct and satisfactory response from the chatbot is received at this stage. The output of the model reflects the probability of unsuccessful completion of all previous stages, as a result of which the process proceeds to the i-th step. The first term of the sum determines the contribution of each possible step of the process completion to the overall average value, weighted by the probability that this particular step will become decisive. The second term takes into account the limiting case when none of the n stages leads to the successful completion of the request processing, and the process is forcibly terminated after passing the entire sequence of conditions.

Thus, model (3) makes it possible to quantitatively assess the average depth of processing a user request, which is an important indicator of the effectiveness of using specialized dictionaries in chatbots with generative artificial intelligence.

The proposed register model (Fig. 1) is expedient to interpret as a formalized representation of the space of possible terms of a specialized XML dictionary used when processing user queries by a chatbot with generative artificial intelligence. In this context, each cell of the register corresponds to a potential combination of characters of length n, formed from the alphabet of the sample q, that is, a possible term or word form. The active cells of the model correspond to the actually existing and described terms in the XML dictionary, for which definitions, semantic attributes, and domain relations are given. The total number of such cells determines the actual volume of the dictionary, while the total number of possible combinations characterizes the full space of potential terms within the selected alphabet and word length.

The small value of r in formula (4) reflects the high specialization of the dictionary, in which only a small part of the possible combinations of symbols corresponds to relevant domain-oriented terms. This, in turn, justifies the need to use semantic search and contextual filtering mechanisms when forming a user query, since the probability of a random match with a correct term is low.

Thus, the parameter r can be considered as a quantitative characteristic of the saturation of a specialized XML dictionary, which is taken into account when building a logical-probabilistic model (Fig. 2, 3). This directly affects the depth and number of stages of query processing by the chat bot (Fig. 4).

The searched active TA cell (Fig. 2) corresponds to the XML dictionary term that is most relevant to the query sample specified by the user. The p_i values assigned to active cells are interpreted as the probabilities that the corresponding term will be selected by the system as a correct search result when accessing the dictionary (Fig. 3). Given the lexicographic ordering of the XML dictionary, the p_i values are distributed along the register randomly, which reflects the lack of a direct correspondence between the position of the term in the XML file and its relevance to a specific query. This justifies the need for a preliminary ranking of active cells in descending order of p_i (Fig. 4), which corresponds to the semantic search and scoring stage in the RAG pipeline.

Fitting the discrete p_i distribution to a continuous function $p(x)$ makes it possible to proceed to a probabilistic assessment of the success of completing the query processing. In this case, the value of $p(x)$ function at point x_0 characterizes the probability that the desired active TA cell will be found before reaching the threshold level of looking up dictionary terms.

Thus, the probability of successful completion of the user query processing in the RAG architecture is directly determined by the integral characteristics of the function $p(x)$, which reflects the density and distribution of relevant terms in the specialized XML dictionary. A decrease in the average number of the process completion step described by model (3) corresponds to an increase in the quality of the dictionary structuring and the efficiency of semantic search mechanisms. This distinguishes our result from [15, 16] due to the clear structuring of the chatbot training process.

Within the practical implementation of the above scheme (Fig. 5), the active TA cells correspond to the <entry> elements of the XML dictionary, and the p_i values are formed based on the semantic proximity of the query to the term, in contrast to [13]. The advantage is due to the fact that the proposed algorithm for forming SSLD is based on the control of user requests, adaptive ranking of terms of the current directory, and step-by-step construction of prompts for query refinement. This approach provides probabilistically controlled narrowing of the search area, reducing the average depth of query processing and improving the quality of response generation in the RAG architecture.

For the above tests 2, 3, 4, 6, 7, 8 ($N \leq 1.1 \cdot 10^4$) the value $D = N$ was adopted. A linear model was used. The keyword coincided with the desired TA. For tests 1, 5 ($N \geq 10^6$), $D = 10^3$ was determined; a logical-probabilistic model was used. The test results (values per session $\overline{v}$, $\overline{t}$, $\overline{m}$ for the corresponding queries) were averaged. As a result, we obtained average signature values of responses with an error of 0.004%. Thus, for tests 2, 3, 4, 6, 7, 8, training was carried out on all positions without errors.

Fig. 6 shows the averaged results of modeling the exponential distribution of queries and the corresponding proba-

bilities p_j used in the process of training a chatbot with step-by-step prompts. Fig. 6, *a* illustrates the initial exponential probability distribution. A characteristic feature is a sharp decline in probabilities with increasing access number *j* and a high concentration of the distribution mass on the first elements. Such a distribution is typical for natural language and information processes. A small number of queries or tokens are used very often, while specialized queries appear infrequently. Fig. 6, *b* shows probabilities after hashing. The result is the disappearance of the monotonic exponential structure, and the p_i values are distributed quasi-randomly over the entire range. In addition, the probabilities become more uniform, without obvious dominant elements.

Special features of our results are as follows:

a) before hashing, the model strongly depends on a limited set of frequent requests. This can lead to overtraining and a decrease in the generalization ability of the chatbot to subject-specific knowledge;

b) after hashing, the influence of individual high-frequency elements decreases, the data balance and stability of learning improve.

The results of this study could be used to adapt chatbots to the specificity of subject areas on project support platforms and in scientific information systems. The approach could also be applied to improve the mechanisms for forming clarifying queries in intelligent dialog systems with subject-specific knowledge.

A limitation of our work is that the effectiveness of SSLD construction directly depends on the relevance, completeness, and structure of the current term reference. In the presence of outdated or incomplete dictionaries, probabilistic ranking may lead to an erroneous narrowing of the search area. The same can be said for the algorithm based on the analysis of real user queries. The presence of errors, incorrect formulations, or ambiguous requests can distort the probability estimates and affect the quality of step-by-step prompts.

A prospect for further research is to investigate the possibilities of building SSLD to train chatbots based on other types of prompts. Another interesting area is the possibility of designing a mechanism for direct training of the chatbot by the user to adapt it to personal needs.

7. Conclusions

1. A model for the process of consecutive independent user requests has been built, which makes it possible to obtain the mathematical expectation of the stage number at which the processing of the user request by the chat bot is completed. This is due to the presence of a check for the correspondence of the request to the terms of a specialized dictionary. The result is the clarification of the semantic context and the generation of a relevant response.

2. An approach to forming a linear model of a specialized dictionary has been proposed, which makes it possible to find a combination of words corresponding to the request by searching through terms. This is due to the fact that in the context each cell of the register corresponds to a potential combination of symbols of a certain length. The result is the comparison of the request term to the term or word form of the dictionary.

3. An approach to building a logical-probabilistic model of a specialized dictionary has been proposed, which makes it possible to choose a target cell. This cell corresponds to the searched word from the specialized dictionary, provided that a certain sample is specified. This is due to the fact that a word is found that corresponds to the term of the XML dictionary and is the most relevant to the query sample specified by the user. The result is a probability distribution of possible word combinations for answering the query.

4. A methodology for constructing a specialized dictionary to train chatbots with step-by-step prompts has been presented. An implementation algorithm and its testing are given. In tests 2–4, 6–8, using a linear model, training was performed correctly for all positions without errors. For tests 1 and 5, a logical-probabilistic model was used with preliminary averaging of session results. This allowed us to obtain average signature values of the response with an error of 0.004% and ensure the stability of the results.

Conflicts of interest

The authors declare that they have no conflicts of interest in relation to the current study, including financial, personal, authorship, or any other, that could affect the study, as well as the results reported in this paper.

Funding

The study was conducted without financial support.

Data availability

All data are available, either in numerical or graphical form, in the main text of the manuscript.

Use of artificial intelligence

The authors used artificial intelligence technologies within acceptable limits to provide their own verified data, which is described in the research methodology section.

Acknowledgments

The authors express their gratitude to employees at the Educational and Methodological Laboratory of Digital Education and Artificial Intelligence, the Hryhorii Skovoroda University in Pereiaslav for their assistance in conducting the experiment and testing the devised approach.

Authors' contributions

Olha Kryazhych: Conceptualization, Methodology, Writing – original draft, Writing – review & editing; **Viacheslav Riznyk**: Project administration, Resources, Validation; **Vasyl Vasenko**: Supervision, Data Curation, Writing – review & editing; **Vasyl Yakuba:** Data Curation, Formal analysis; **Kateryna Iushchenko**: Investigation, Formal analysis; **Oleksii Kuprin**: Software; **Oleksandr Tsyrul**: Software

References

1. Yang, C., Zhao, R., Liu, Y., Jiang, L. (2025). Survey of specialized large language model. arXiv. https://arxiv.org/abs/2508.19667
2. Adavala, K. M., Adavala, O. (2025). Domain-specific knowledge and context in large language models: challenges, concerns, and solutions. IAES International Journal of Artificial Intelligence (IJ-AI), 14 (4), 2568. https://doi.org/10.11591/ijai.v14.i4.pp2568-2578
3. Zhu, Y., Yuan, H., Wang, S., Liu, J., Liu, W., Deng, C. et al. (2025). Large Language Models for Information Retrieval: A Survey. ACM Transactions on Information Systems, 44 (1), 1–54. https://doi.org/10.1145/3748304
4. Ai, Q., Bai, T., Cao, Z., Chang, Y., Chen, J., Chen, Z. et al. (2023). Information Retrieval meets Large Language Models: A strategic report from Chinese IR community. AI Open, 4, 80–90. https://doi.org/10.1016/j.aiopen.2023.08.001
5. Sharma, K., Kumar, P., Li, Y. (2025). OG-RAG: Ontology-grounded retrieval-augmented generation for large language models. Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, 32950–32969. https://doi.org/10.18653/v1/2025.emnlp-main.1674
6. Manda, P. (2025). Large Language Models in Bio-Ontology Research: A Review. Bioengineering, 12 (11), 1260. https://doi.org/10.3390/bioengineering12111260
7. Barron, R. C., Grantcharov, V., Wanna, S., Eren, M. E., Bhattarai, M., Solovyev, N. et al. (2024). Domain-Specific Retrieval-Augmented Generation Using Vector Stores, Knowledge Graphs, and Tensor Factorization. 2024 International Conference on Machine Learning and Applications (ICMLA), 1669–1676. https://doi.org/10.1109/icmla61862.2024.00258
8. Fareedi, A. A., Ismail, M., Ahmed, S., Gagnon, S., Ghazawneh, A., Arooj, Z., Nazir, H. (2025). Enriching Human–AI Collaboration: The Ontological Service Framework Leveraging Large Language Models for Value Creation in Conversational AI. Knowledge, 6 (1), 2. https://doi.org/10.3390/knowledge6010002
9. Mukanova, A., Nazyrova, A., Zulkhazhav, A., Lamasheva, Z., Dauletkaliyeva, A. (2025). Development of an Intelligent Information Retrieval System Based on Ontology, Linguistic Algorithms and Large Language Models. Applied Sciences, 15 (22), 12271. https://doi.org/10.3390/app152212271
10. Ahmad, J. M., Liu, Y., Kim, J.-D., Yao, X., Larmande, P., Xia, J. (2025). A curation system of rice trait ontology with reliable interoperation by LLM and PubAnnotation. Genomics & Informatics, 23 (1). https://doi.org/10.1186/s44342-025-00058-z
11. Chen, L.-C., Pardeshi, M. S., Liao, Y.-X., Pai, K.-C. (2025). Application of retrieval-augmented generation for interactive industrial knowledge management via a large language model. Computer Standards & Interfaces, 94, 103995. https://doi.org/10.1016/j.csi.2025.103995
12. Wen, J., Liu, D., Xie, Y., Ren, Y., Wang, J., Xia, Y., Zhu, P. (2025). AcuGPT-Agent: An LLM-powered intelligent system for acupuncture-based infertility treatment. Neurocomputing, 652, 131116. https://doi.org/10.1016/j.neucom.2025.131116
13. Rodríguez-Muñoz-de-Baena, I., Coronado-Vaca, M., Vaquero-Lafuente, E. (2025). Fine-tuning transformer models for M&A target prediction in the U.S. ENERGY sector. Cogent Business & Management, 12 (1). https://doi.org/10.1080/23311975.2025.2487219
14. Byrd, C., Kingsbury, C., Niell, B., Funaro, K., Bhatt, A., Weinfurtner, R. J., Ataya, D. (2025). Appropriateness of acute breast symptom recommendations provided by ChatGPT. Clinical Imaging, 125, 110549. https://doi.org/10.1016/j.clinimag.2025.110549
15. Brown, E. D. L., Ward, M., Maity, A., Mittler, M. A., Larry Lo, S.-F., D'Amico, R. S. (2024). Enhancing Diagnostic Support for Chiari Malformation and Syringomyelia: A Comparative Study of Contextualized ChatGPT Models. World Neurosurgery, 189, e86–e107. https://doi.org/10.1016/j.wneu.2024.05.172
16. Ni, W., Shen, Q., Liu, T., Zeng, Q., Xu, L. (2023). Generating textual emergency plans for unconventional emergencies – A natural language processing approach. Safety Science, 160, 106047. https://doi.org/10.1016/j.ssci.2022.106047
17. Ganzinger, M., Kunz, N., Fuchs, P., Lyu, C. K., Loos, M., Dugas, M., Pausch, T. M. (2025). Automated generation of discharge summaries: leveraging large language models with clinical data. Scientific Reports, 15 (1). https://doi.org/10.1038/s41598-025-01618-7
18. Xu, Y., Wang, T., Yuan, Y., Huang, Z., Chen, X., Zhang, B. et al. (2025). LLM-Enhanced Framework for Building Domain-Specific Lexicon for Urban Power Grid Design. Applied Sciences, 15 (8), 4134. https://doi.org/10.3390/app15084134
19. Keng-Jung, P., Chin-Hung, K., Cheng-Yen, W., Peng, J.-W., Huang, C.-Y., Chen, J.-C. (2021). Analyze the subordination structure between domain-specific vocabulary and meaning with the Word2Vec training process. 2021 IEEE International Conference on Consumer Electronics-Taiwan (ICCE-TW), 1–2. https://doi.org/10.1109/icce-tw52618.2021.9602966
20. Xu, K., Feng, Y., Li, Q., Dong, Z., Wei, J. (2025). Survey on terminology extraction from texts. Journal of Big Data, 12 (1). https://doi.org/10.1186/s40537-025-01077-x
21. Lu, R.-S., Lin, C.-C., Tsao, H.-Y. (2024). Empowering Large Language Models to Leverage Domain-Specific Knowledge in E-Learning. Applied Sciences, 14 (12), 5264. https://doi.org/10.3390/app14125264
22. Kryazhych, O., Ivanov, I., Iushchenko, K., Kupri, O., Vasenko, O., Riznyk, V., Ryzhkov, O. (2025). Devising an approach to preventing information chaos in chat bots using generative artificial intelligence. Eastern-European Journal of Enterprise Technologies, 2 (2 (134)), 84–95. https://doi.org/10.15587/1729-4061.2025.324957