# Bibliometric analysis and experimental assessment of chatbot training approaches

Roman O. Liashenko[1], Serhiy O. Semerikov[1,2,3,4,5]

[1]*Kryvyi Rih State Pedagogical University, 54 Universytetskyi Ave., Kryvyi Rih, 50086, Ukraine*
[2]*Institute for Digitalisation of Education of the NAES of Ukraine, 9 M. Berlynskoho Str., Kyiv, 04060, Ukraine*
[3]*Zhytomyr Polytechnic State University, 103 Chudnivsyka Str., Zhytomyr, 10005, Ukraine*
[4]*Kryvyi Rih National University, 11 Vitalii Matusevych Str., Kryvyi Rih, 50027, Ukraine*
[5]*Academy of Cognitive and Natural Sciences, 54 Universytetskyi Ave., Kryvyi Rih, 50086, Ukraine*

**Abstract**
The rapid development of natural language processing and artificial intelligence has led to growing interest in creating chatbots capable of engaging in human-like dialogue. This study conducts a bibliometric analysis of research on chatbot training to identify key concepts, trends, and promising directions. The analysis of 549 publications from the Scopus database using VOSviewer reveals four main clusters of research: natural language processing techniques, application of NLP technologies in society, use of machine learning for NLP, and chatbots in education and service sectors. To experimentally evaluate chatbot training approaches, two datasets are created from scientific publications and used to fine-tune GPT-2 language models. The results demonstrate the feasibility and effectiveness of transfer learning for adapting pre-trained models to domain-specific data. This research provides insights into the state-of-the-art in chatbot development and highlights opportunities for future work on creating specialized conversational agents.

**Keywords**
chatbot, natural language processing, bibliometric analysis, transfer learning, GPT-2

## 1. Introduction

The rapid development of artificial intelligence (AI) and natural language processing (NLP) technologies has sparked a surge of interest in creating software agents capable of engaging in human-like dialogue, commonly known as chatbots. Leading technology companies such as Google [1], Microsoft [2], Meta (Facebook) [3], and OpenAI [4] are heavily investing in chatbot development. The success of projects like OpenAI's ChatGPT [5] highlights the immense potential for deploying such systems across various domains of human activity. Recent research on chatbots predominantly focuses on two main areas: (1) utilizing chatbots for student education and learning [6]; and (2) developing question-answering systems that train their own models using user-provided data [7]. However, the effective training of chatbots also raises critical scientific and practical challenges related to system reliability, safety, and ethics.

As AI systems capable of understanding and generating natural language, chatbots have the potential to revolutionize various aspects of human activities [8]. They can automate routine tasks, provide intelligent user support, enable personalized learning experiences, and much more. Breakthroughs in chatbot development could fundamentally transform how humans interact with computer systems, ultimately boosting productivity and enhancing learning outcomes. Nevertheless, addressing the ethical and security implications of chatbot technologies is crucial to ensure their responsible deployment for the benefit of society.

This study aims to advance the field of chatbot development by conducting a comprehensive bibliometric analysis of research on chatbot training and experimentally evaluating key training approaches.

*CS&SE@SW 2024: 7th Workshop for Young Scientists in Computer Science & Software Engineering, December 27, 2024, Kryvyi Rih, Ukraine*
romanslss777@gmail.com (R. O. Liashenko); semerikov@gmail.com (S. O. Semerikov)
https://acnsci.org/semerikov (S. O. Semerikov)
0009-0000-2614-6997 (R. O. Liashenko); 0000-0003-0789-0272 (S. O. Semerikov)

The bibliometric analysis will identify trends, key concepts, and knowledge gaps in the existing literature, providing valuable insights to guide future research efforts. The experimental assessment will involve fine-tuning state-of-the-art language models on domain-specific datasets to create specialized chatbots. The findings will demonstrate best practices and practical strategies for adapting chatbots to target domains.

The main **objectives** of this research are threefold:

1. Conduct a bibliometric analysis of publications on chatbot training to identify key research concepts, trends, and promising directions for future work.
2. Compare the performance of popular chatbot training models and establish evaluation criteria for assessing their effectiveness.
3. Develop two domain-specific datasets and experimentally evaluate the performance of fine-tuned large language models on these datasets.

To achieve these objectives, we employ a multi-method approach. First, we perform *bibliometric analysis* on search results from the Scopus database to: (a) identify the chronological boundary marking a steady increase in chatbot training publications; (b) construct a map of keyword co-occurrences; (c) cluster keywords into thematic groups; and (d) determine the central research concepts (as decribed in our previous works [9, 10]). Second, we use *state-of-the-art language models* – namely GPT-3.5, GPT-4.0, Google Bard, and Claude 2 – to generate cluster names and descriptions based on the keyword analysis. Finally, we apply *software engineering methods*, including system design, implementation, testing, and experimental trials, to fine-tune the language models and evaluate their performance.

The results of this study have both theoretical and practical implications. The bibliometric analysis provides a systematic mapping of the chatbot training research landscape, identifying key themes, trends, and research gaps. This can guide researchers in identifying promising avenues for future work and inform funding agencies about areas needing further investigation. The experimental findings offer organizations actionable insights into effective methods for creating domain-specific chatbots using transfer learning techniques. This can help businesses harness the power of conversational AI to enhance their operations and customer experiences.

The remainder of this paper is structured as follows. Section 2 presents the bibliometric analysis, describing the methodology and discussing the results. Section 3 reviews the main approaches to chatbot training, including supervised learning, reinforcement learning, and transfer learning, and outlines evaluation metrics. Section 4 details the experimental procedure, dataset creation, model selection, and fine-tuning process. Finally, section 5 concludes the paper by summarizing the key findings and outlining directions for future research.

## 2. Bibliometric analysis

### 2.1. Rationale

Both scientists and IT company developers are actively working in the field of creating and training chatbots. Certain aspects of this issue are covered in a number of publications. In particular, the search results in the DeepLearning.AI blog [11] for 2019-2023 provide an opportunity to highlight the following practically solved tasks:

- since 2020, chatbots (Generative BST from Facebook and Google Meena) can be used for *short dialogues on general topics* [12, 13];
- since 2023, Microsoft has provided the ability to use chatbots in Office 365 and Windows to *boost productivity* [14];
- since 2020, chatbots are actively *used in business* for customer service, sales, etc. [15];
- since 2023, Google and Microsoft have been introducing chatbots into *search engines* [16].

DeepLearning.AI blog articles also point to knowledge gaps and open problems that exist in this field:

- medical chatbots can make false diagnoses [17];
- chatbots such as BlenderBot 3 and Galactica can exhibit bias, toxicity, and fact distortion [18];
- high cost of deploying large language models [19];
- lack of transparency in chatbot models can raise ethical questions [20].

An analysis of recent research and publications reveals the need to consolidate efforts to determine relevant areas and priorities for research in order to effectively train chatbots.

## 2.2. Results

For bibliometric analysis, the Scopus bibliographic database published by Elsevier was chosen as one of the largest abstract databases, indexing more than 42,000 periodicals and containing mostly high-quality scientific sources in technical, natural, medical and social sciences, which makes it representative for analysing publications on the problem of chatbot training at the intersection of several scientific fields. The balanced coverage of various scientific disciplines in Scopus provides the necessary completeness to identify key trends and research directions in the chosen field.

To identify in which works the key research concepts of “training” and “chatbots” occur, a search was performed on November 18, 2023 in the Scopus database [21]. Those documents were selected in the titles, abstracts and keywords of which the words “chatbot” and “training” occurred simultaneously (figure 1).

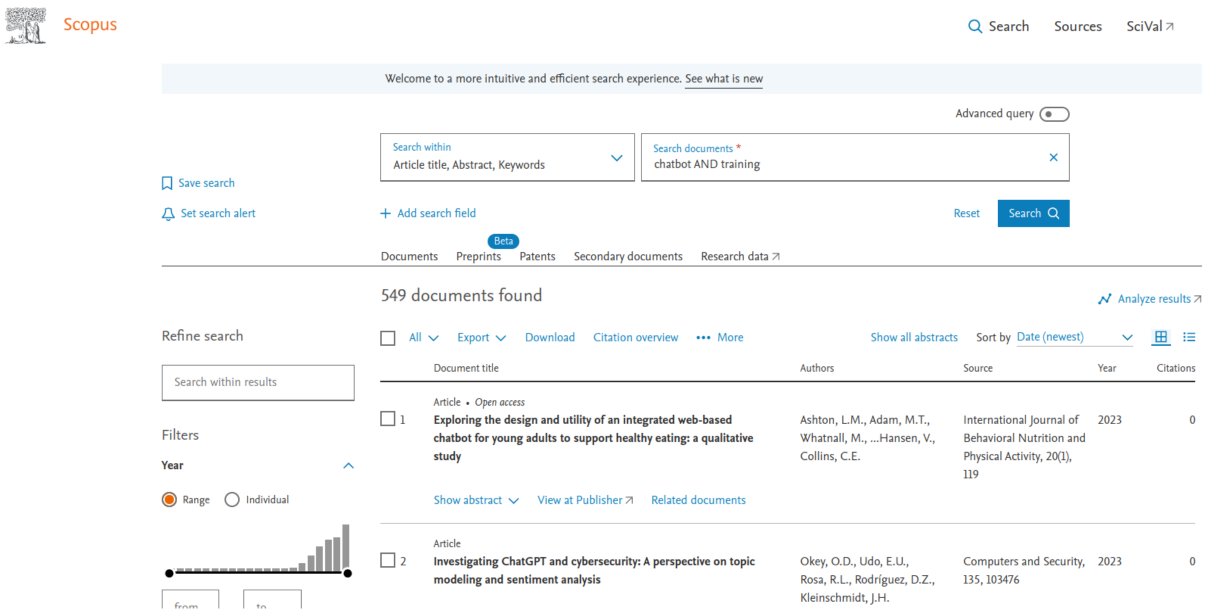


**Figure 1:** Search results in the Scopus database for the words “chatbot”, “training”.

The total number of documents – 549; the distribution of documents by year is shown in figure 2.

A review of the keywords shows that the most frequently occurring ones are: Chatbot (244 times), Chatbots (221), Artificial Intelligence (126), Natural Language Processing Systems (90), Natural Language Processing (90). Sorting the keywords alphabetically provides an opportunity to determine which of them differ only in number (singular and plural) and to convert all to singular:

- Chatbots is the plural of Chatbot;
- Conversational Agents is the plural of Conversational Agent;
- Convolutional Neural Networks is the plural of Convolutional Neural Network;
- Customer Services is the plural of Customer Service (Customer-service);

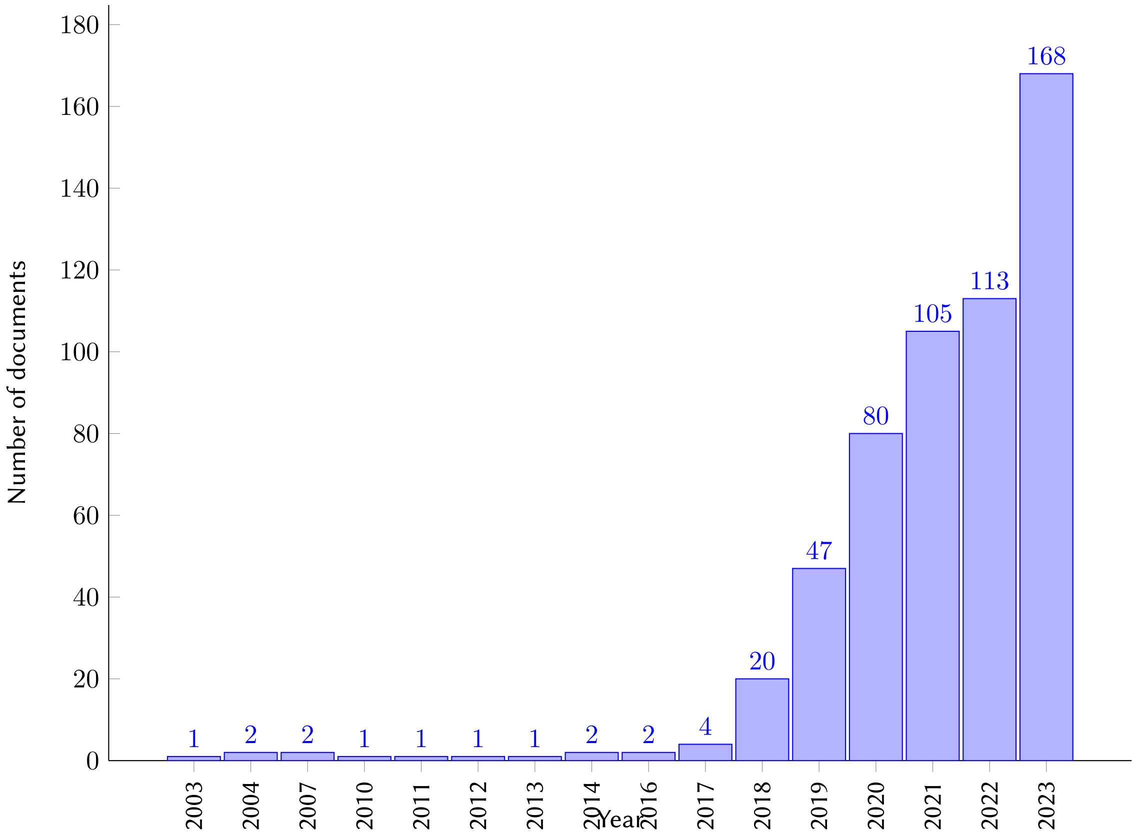


**Figure 2:** Distribution of search results by year.

- Dialogue Generations is the plural of Dialogue Generation;
- Humans is the plural of Human;
- Large Language Models is the plural of Large Language Model;
- LSTM is the abbreviation for Long Short-term Memory;
- Machine-learning is the synonym for Machine Learning;
- Mobile Applications is the plural of Mobile Application;
- NLP is the abbreviation for Natural Language Processing (NAtural Language Processing);
- NLU is the abbreviation for Natural Language Understanding;
- Reinforcement Learnings is the plural of Reinforcement Learning;
- Virtual Assistants is the plural of Virtual Assistant.

The found documents were exported in CSV [22, p. 30] and BibTeX formats. In the CSV file, replacement of plural keywords and abbreviations with singular keywords was performed.

To perform bibliometric analysis, VOSviewer 1.6.20 [23] was used: *Create...* → *Create a map based on bibliographic data* → *Read data from bibliographic database file* → *Scopus* (figure 3).

The map is created with the following parameters: analysis type – by co-occurrence of the term; unit of analysis – all keywords; counting method – full counting.

The total number of keywords – 3705 – can be reduced to 58 by discarding those that occur less than 12 times. From the selected words, the word “article” was additionally excluded as not being specific to the search query.

The constructed map (table 1) is presented in figure 4.

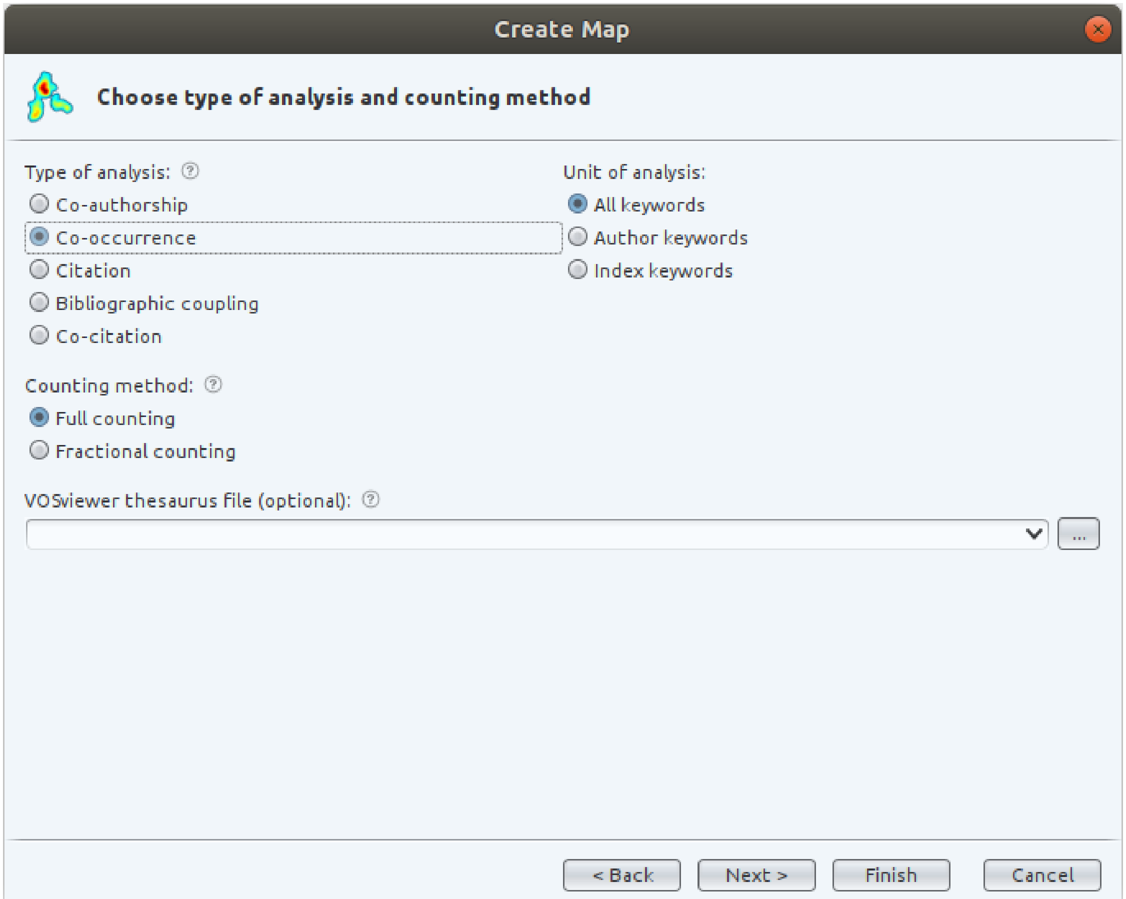


**Figure 3:** Creating a map.

Table 1: Distribution of keywords by clusters.

| Keyword | Cluster | $W_{Links}$ | $W_{Total\ link\ strength}$ | $W_{Occurrences}$ | $S_{Avg.\ pub.\ year}$ | $S_{Avg.\ citations}$ | $S_{Avg.\ norm.\ citations}$ | Significance |
|---|---|---|---|---|---|---|---|---|
| classification (of information) | 1 | 37 | 122 | 24 | 2021.2083 | 3.7917 | 0.5908 | 3 |
| training data | 1 | 36 | 120 | 27 | 2021.1111 | 12.8519 | 1.8358 | 3 |
| speech processing | 1 | 36 | 102 | 26 | 2020.4615 | 9.4231 | 1.1619 | 1 |
| performance | 1 | 36 | 88 | 24 | 2021.5833 | 6.2917 | 1.02 | 2 |
| semantics | 1 | 34 | 90 | 18 | 2020.7222 | 8.8333 | 0.9379 | |
| dialogue systems | 1 | 33 | 96 | 23 | 2020.4348 | 8.4348 | 1.0096 | |
| natural language understanding | 1 | 32 | 123 | 32 | 2020.9375 | 6.0938 | 0.8055 | 3 |
| computational linguistics | 1 | 32 | 114 | 33 | 2020.1515 | 16.1515 | 2.0909 | 6 |
| text processing | 1 | 31 | 97 | 16 | 2021.25 | 4.4375 | 0.6438 | |
| long short-term memory | 1 | 30 | 113 | 24 | 2020.7917 | 4.8333 | 0.6383 | |
| state of the art | 1 | 29 | 54 | 14 | 2021.0714 | 4.0714 | 0.5852 | |
| language model | 1 | 28 | 61 | 14 | 2021.9286 | 3.5 | 0.6501 | 2 |
| question answering | 1 | 28 | 60 | 12 | 2021.1667 | 1.4167 | 0.2476 | |
| information retrieval | 1 | 27 | 51 | 12 | 2020.4167 | 9.3333 | 0.9633 | |

*Continued on next page*

Table 1 – continued from previous page

| Keyword | Cluster | $W_{Links}$ | $W_{Total\ link\ strength}$ | $W_{Occurrences}$ | $S_{Avg.\ pub.\ year}$ | $S_{Avg.\ citations}$ | $S_{Avg.\ norm.\ citations}$ | Significance |
|---|---|---|---|---|---|---|---|---|
| embeddings | 1 | 24 | 53 | 13 | 2020.8462 | 7.5385 | 0.6933 | |
| reinforcement learning | 1 | 23 | 52 | 15 | 2020.0667 | 9.0667 | 1.1396 | |
| response generation | 1 | 17 | 37 | 12 | 2020.75 | 11.5833 | 1.343 | |
| artificial intelligence | 2 | 52 | 452 | 126 | 2021.3333 | 9.8571 | 1.8428 | 6 |
| conversational agent | 2 | 47 | 203 | 50 | 2020.18 | 18.68 | 1.9112 | 1 |
| human | 2 | 37 | 266 | 60 | 2022.1167 | 15.7667 | 2.7326 | 2 |
| adult | 2 | 29 | 117 | 17 | 2021.9412 | 17 | 2.3017 | |
| training | 2 | 29 | 84 | 22 | 2021.7727 | 16.7273 | 3.5069 | 2 |
| male | 2 | 27 | 94 | 13 | 2021.4615 | 23.3846 | 3.0896 | 1 |
| medical education | 2 | 27 | 66 | 12 | 2022 | 3.5833 | 0.557 | |
| female | 2 | 26 | 101 | 16 | 2021.375 | 22.6875 | 2.6942 | |
| controlled study | 2 | 26 | 76 | 15 | 2022.2 | 9.2 | 1.6989 | 1 |
| health care | 2 | 26 | 67 | 15 | 2021.4 | 3.6 | 0.51 | |
| mental health | 2 | 25 | 56 | 13 | 2021.3846 | 22.6154 | 2.2811 | |
| chatgpt | 2 | 24 | 124 | 35 | 2023 | 4.8 | 1.4989 | 2 |
| education | 2 | 24 | 77 | 19 | 2021.1053 | 12.4211 | 3.3792 | 1 |
| covid-19 | 2 | 23 | 47 | 13 | 2021.9231 | 3.5385 | 0.884 | |
| review | 2 | 19 | 70 | 12 | 2021.9167 | 34.25 | 3.3067 | 2 |
| large language model | 2 | 19 | 50 | 12 | 2023 | 2.9167 | 0.9108 | 2 |
| natural language processing | 3 | 54 | 505 | 100 | 2021.28 | 6.29 | 1.1837 | 6 |
| natural language processing systems | 3 | 52 | 484 | 90 | 2020.7667 | 5.5667 | 0.6711 | 3 |
| machine learning | 3 | 50 | 276 | 51 | 2021.4314 | 8.9412 | 1.0081 | |
| deep learning | 3 | 46 | 265 | 56 | 2021.1429 | 5.6607 | 0.6799 | |
| natural languages | 3 | 42 | 235 | 40 | 2021.5 | 4.1 | 0.6042 | |
| language processing | 3 | 37 | 165 | 25 | 2022.56 | 1.6 | 0.5094 | 2 |
| learning algorithms | 3 | 37 | 153 | 26 | 2021.0769 | 3.4231 | 0.4283 | |
| virtual assistant | 3 | 35 | 67 | 15 | 2021.0667 | 13.3333 | 1.5093 | 3 |
| user interfaces | 3 | 29 | 61 | 14 | 2020.8571 | 8.9286 | 1.7491 | 2 |
| convolutional neural network | 3 | 26 | 53 | 12 | 2020.9167 | 6.4167 | 0.7242 | |
| query processing | 3 | 24 | 82 | 14 | 2021.1429 | 1.4286 | 0.3373 | |
| diagnosis | 3 | 24 | 53 | 13 | 2021.5385 | 9.8462 | 1.3824 | 2 |
| chatbot | 4 | 56 | 1065 | 365 | 2021.137 | 6.3014 | 0.901 | 6 |
| learning systems | 4 | 46 | 223 | 43 | 2020.2326 | 7.186 | 0.4899 | 3 |
| students | 4 | 42 | 176 | 38 | 2020.9737 | 8.2105 | 0.9839 | |
| e-learning | 4 | 35 | 128 | 37 | 2020.8378 | 5.0541 | 0.6268 | |
| learn+ | 4 | 27 | 53 | 13 | 2021.7692 | 3.9231 | 0.6622 | 2 |
| sales | 4 | 26 | 64 | 17 | 2021 | 5.5882 | 0.6926 | |
| human computer interaction | 4 | 26 | 53 | 14 | 2019.8571 | 6.0714 | 0.5181 | |
| knowledge based systems | 4 | 26 | 51 | 13 | 2019.5385 | 15.5385 | 0.7796 | 2 |
| personnel training | 4 | 25 | 62 | 18 | 2020.8889 | 6.5556 | 1.4229 | 1 |
| curricula | 4 | 24 | 58 | 12 | 2021 | 9.5833 | 2.41 | 3 |
| customer service | 4 | 24 | 49 | 15 | 2021.2 | 4.2667 | 0.4819 | 1 |
| engineering education | 4 | 24 | 46 | 12 | 2020.75 | 4.5833 | 0.4115 | |

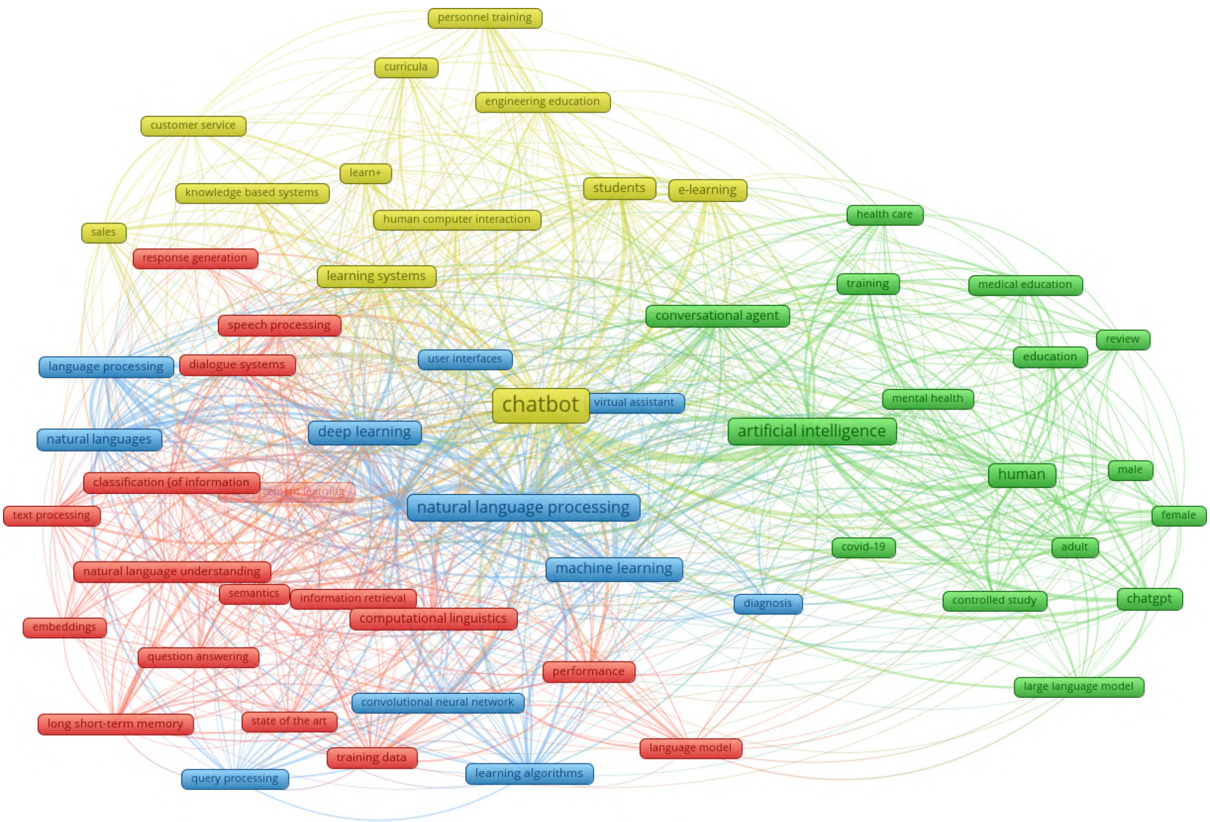


**Figure 4:** Map of keyword connections.

In table 1 the following notations are used [22, pp. 6, 38]:

- keyword – a term characteristic of a publication, defined by the author or indexing system;
- cluster – the number of the keyword group;
- $W_{Links}$ – the number of links of a keyword with other keywords;
- $W_{Total\ link\ strength}$ – the total strength of the links of a keyword with other keywords (here the number of publications in which two terms co-occur);
- $W_{Occurrences}$ – the number of documents in which a keyword occurs;
- $S_{Avg.\ pub.\ year}$ – the average publication year of the documents in which a keyword occurs;
- $S_{Avg.citations}$ – the average number of citations received by the documents in which a keyword occurs;
- $S_{Avg.\ norm.\ citations}$ – the average normalized number of citations received by the documents in which a keyword occurs;
- significance – calculated as the sum, where for each keyword, 2 is selected if it has the highest value in the column, and 1 if it has the value preceding the highest.

In appendix A, the results of applying generative chatbots ChatGPT (GPT 3.5), Bing (GPT 4.0, 3 modes), Claude 2 and Google Bard to determine the names of clusters are presented (table 2).

The engaged large language models generated several variants of names and descriptions of clusters based on the analysis of keywords that are part of their composition. For each cluster, all variants proposed by the models are provided with notations A.1, A.2, and so on.

The choice of the final names of clusters, presented in the "Chosen name" column of table 2, was carried out through careful analysis and critical evaluation by the authors of all suggestions generated

**Table 2**
Names of keyword clusters.

| No. | Suggestions from chatbots | Chosen name |
|---|---|---|
| 1 | Advanced Language Processing and AI Applications (A.1); Language Processing and Information Retrieval (A.2); Natural Language Processing (NLP) (A.3); Natural Language Understanding and Generation (A.4); Natural Language Processing Applications (A.5); Technical NLP (A.6) | Natural Language Processing (A.3) |
| 2 | AI in Healthcare and Social Context (A.1); AI in Healthcare and Education (A.2); Artificial Intelligence (A.3); Artificial Intelligence and Health Care (A.4); Conversational AI and Impacts (A.5); Applications of NLP (A.6) | Application of Natural Language Processing Technologies in Society (A.6, A.5, A.1) |
| 3 | Core Technologies in Natural Language Processing and Machine Learning (A.1); Machine Learning and User Interaction (A.2); Machine Learning (A.3); Natural Language Processing and Machine Learning: (A.4); Core AI and ML Techniques (A.5); Machine Learning and Deep Learning for NLP (A.6) | Application of Machine Learning for Natural Language Processing (A.6, A.4) |
| 4 | Educational and Service-oriented AI Applications (A.1); AI in Education and Customer Service (A.2); Learning Systems (LS) (A.3); Chatbot and Education (A.4); AI for Training and Customer Service (A.5); Chatbots and Learning Systems (A.6) | Chatbots in Education and Service Sector (A.1, A.2, A.4, A.5) |

by different models. The key selection criteria were meaningfulness, clarity, brevity of names with simultaneous maximum coverage of key terms in the composition of clusters.

The chosen cluster names not only generalize their content but also logically reflect the main research directions in the field of chatbot training identified by the results of bibliometric analysis: 1) basic methods of natural language processing; 2) application of relevant technologies in various spheres of social activity; 3) the use of machine learning as a leading toolkit for developing natural language processing systems; 4) the range of key applications of chatbots, in particular in the educational sector and service sector.

Thus, the final choice of cluster names was the result of combining the capabilities of large language models with the expert opinion of the authors and is based on clear justified positions aimed at avoiding ambiguities in the interpretation of the obtained results.

Figure 5 shows a map of connections of keywords with significance not less than 1, i.e. the most significant keywords belonging to the four identified clusters. This map allows analysing the relationships between the central concepts of the study and visualizing the leading directions within each cluster.

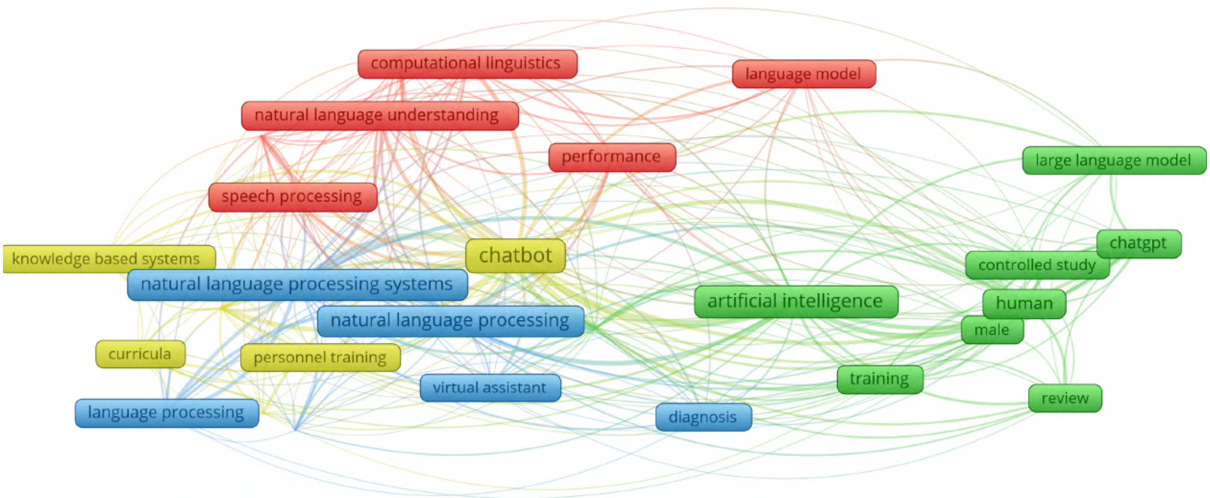


**Figure 5:** Map of connections of the most significant keywords.

The construction of this focused map was carried out with the aim of concentrating attention on the most important keywords, selected on the principle of maximum and pre-maximum values of a number of indicators (number of links, total weight of links, number of mentions in publications, average values of publication year, citation, etc.). The visualization of the most significant terms and their relationships allows identifying priority areas in each of the studied fields to outline promising directions for future scientific research.

Such an approach allows detailing and deepening the conclusions made based on the general keyword map, focusing directly on the central concepts of chatbot training as the leading artificial intelligence agents.

# 3. Chatbot training models

## 3.1. Supervised learning

Supervised learning is one of the main approaches to machine learning, which is widely used to build chatbots and other dialogue systems [24, p. 3940] (figure 6). This approach involves training the model on labeled data, where each example contains input information (user query) and the corresponding target response (chatbot reply).

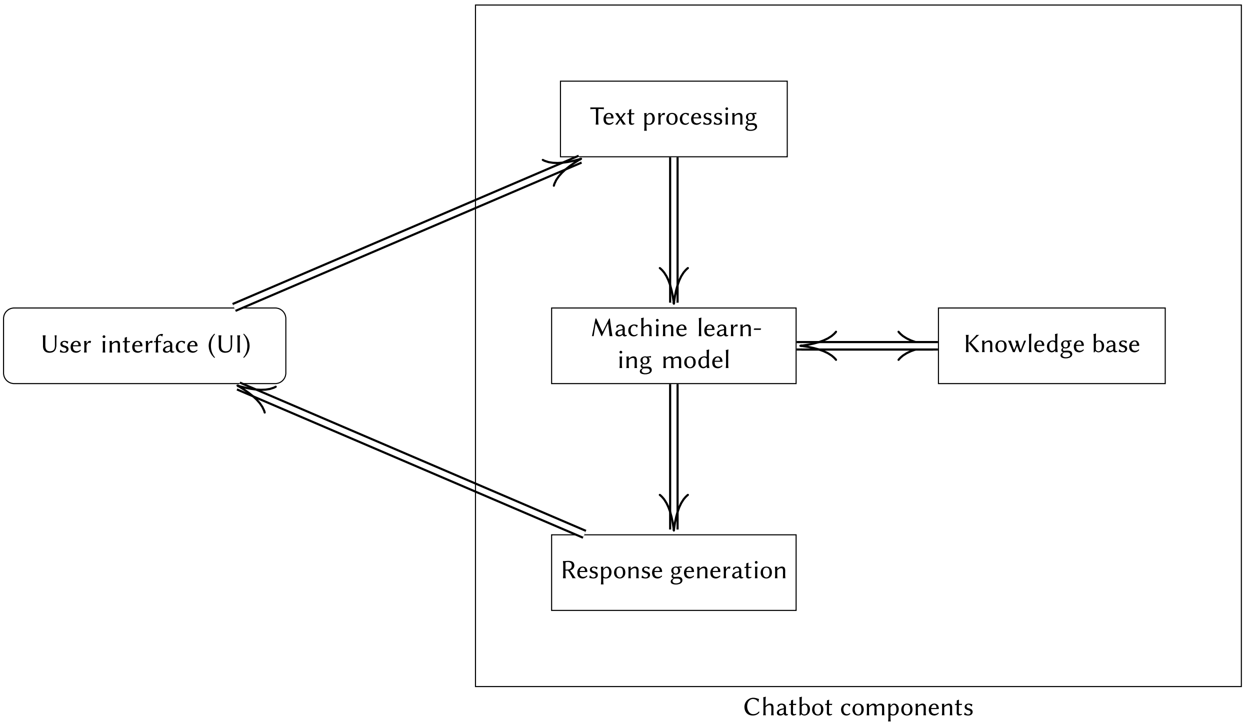


**Figure 6:** General scheme of chatbot operation (based on [25, p. 2]).

Patil et al. [25] distinguish the following chatbot components [25, p. 2]):

- *text processing* – word embeddings are vector representations of words within a specific vocabulary, allowing for better implementation and utilization of statistical machine learning models;
- *machine learning model* – the concept of an artificial neural network, which is widely used for input processing, classification and generation of the most appropriate response to the input query;
- *knowledge base* – the dataset used for training the model, which can be open or closed in subject area (domain): open domain chatbots are inferior in the relevance and accuracy of responses, while closed domain chatbots perform well due to the limited but clearly defined scope of the dataset;

- *response generation*: the response returned to the input query is either selected from a base or generated depending on the features of input vectors, dictionary and trained classifier.

The supervised learning process for chatbots includes the following steps:

1. Selection of the model architecture.
2. Collection and annotation of training data.
3. Data preprocessing.
4. Model training.
5. Evaluation of the quality of the trained model.

Among the challenges of supervised learning, the need for a large volume of high-quality labeled data can be highlighted, the collection and annotation of which can be a laborious and costly process [24, p. 3940]. One way to address this is to use semi-supervised learning, which allows training the model on both labeled and unlabeled data [24, p. 3943].

#### 3.1.1. Seq2Seq models based on recurrent neural networks

Seq2Seq (Sequence-to-Sequence) models based on recurrent neural networks with the application of LSTM (Long Short-Term Memory) modules are one of the approaches to building dialogue systems using machine learning methods [26, p. 230].

The Seq2Seq architecture consists of two main components: an encoder and a decoder. The encoder processes the input sequence (user query) and generates its vector representation in the hidden space. The decoder, in turn, uses this representation to generate the output sequence (chatbot response) [25, p. 9].

A key feature of Seq2Seq models is the use of recurrent layers, particularly LSTM, for processing sequential data [25, p. 4]. LSTM modules allow effectively modeling long-term dependencies in sequences, which is critically important for generating coherent and contextually relevant responses in dialogue [25, p. 5].

Seq2Seq models based on LSTM have certain limitations. In particular, they can suffer from the problem of "vanishing gradients" when processing long sequences [25, p. 4].

Nevertheless, LSTM-based Seq2Seq architectures still remain an important building block of modern dialogue systems. They are often used as base models, which can then be improved and extended through other methods such as attention modules or hierarchical architectures [24, p. 3941-3942].

In the context of chatbot development, Seq2Seq models based on LSTM have proven to be an effective tool for generating responses to user queries in various subject areas [27].

#### 3.1.2. Transformer architectures

Transformer architectures, particularly models of the GPT (Generative Pre-trained Transformer) family, represent a state-of-the-art approach to building language models and dialogue systems. These architectures have gained extraordinary popularity due to the fact that they consistently outperform other language models such as recurrent neural networks. These models are based on the self-attention mechanism and are capable of efficiently processing and generating sequences of arbitrary length, storing longer conversation histories, which leads to consistent, contextual, and improved dialogue [28, p. 2].

The transformer architecture consists of several encoder and decoder blocks, each containing self-attention and feed-forward layers (figure 7). This allows transformers to process all elements of the input sequence in parallel [29, p. 2], which significantly speeds up the learning process compared to recurrent models [29, p. 10].

Researchers have also demonstrated how the performance of large language models trained on a large corpus of data can be improved by fine-tuning on specific tasks. This can be clearly observed when we look at the GPT series from OpenAI (figure 8): GPT, GPT-2, GPT-3, and GPT-4, which are

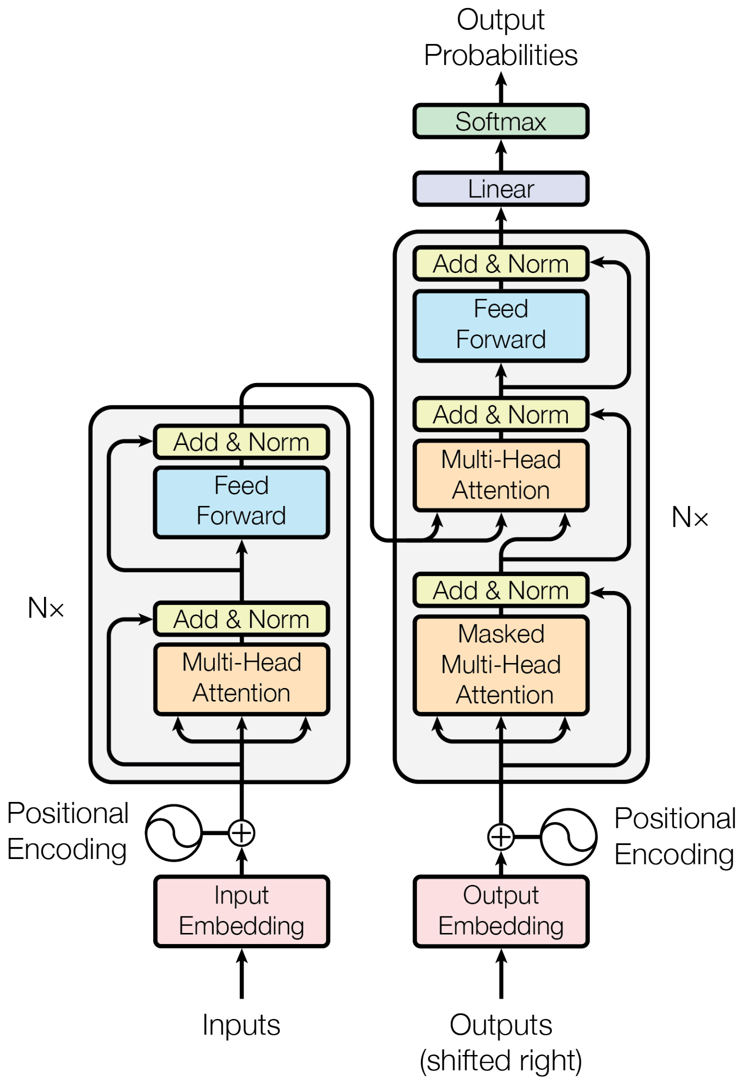


**Figure 7:** The transformer architecture [29, p. 3].

capable of performing tasks such as question answering, reading comprehension, text summarization, text generation, or conversation modeling [28, p. 2].

Despite the impressive performance of GPT models, it is known that they exhibit a phenomenon called hallucination, where they generate results that are contextually implausible or incompatible with the real world [30, p. 1]. Despite this, transformer architectures, and particularly GPT models, have become the de facto standard for building modern dialogue systems and chatbots. Their ability to process context, generate human-like responses, and adapt to different subject areas makes them a powerful tool for creating intelligent assistants and virtual conversationalists.

### 3.2. Reinforcement learning

Reinforcement learning is one of the approaches to machine learning that is gaining popularity in the field of building chatbots and other dialogue systems. Unlike supervised learning, where the model is trained on labeled pairs of "query-response", reinforcement learning allows the model to learn through interaction with the environment and receiving feedback in the form of rewards for its actions.

For the case where the environment is the user, the reinforcement learning process for chatbots can be represented as follows:

1. The agent (chatbot) interacts with the user at discrete points in time.

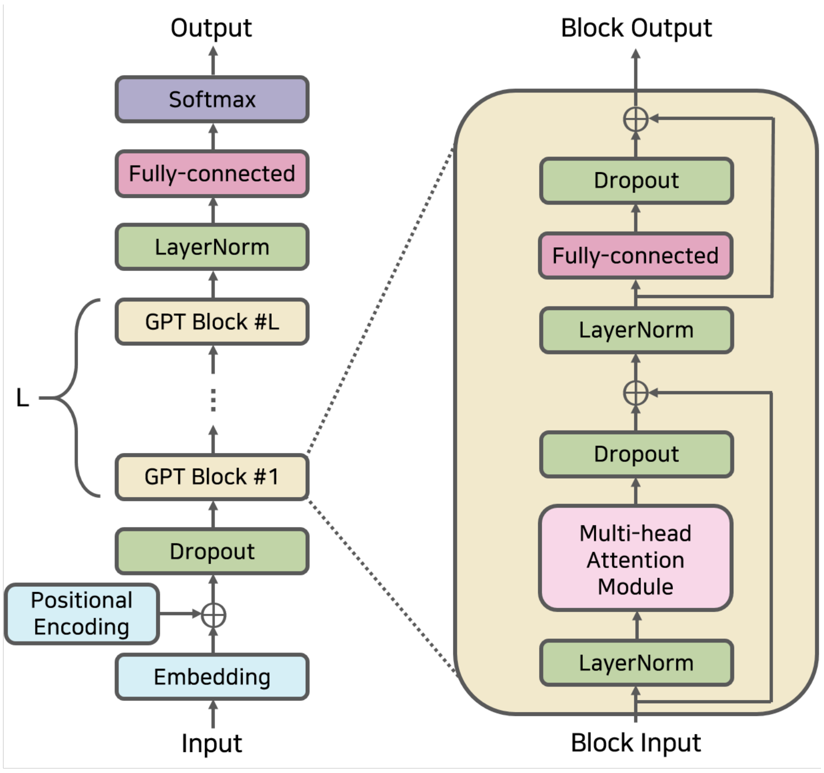


**Figure 8:** Conceptual architecture of a GPT model [30, p. 4].

2. At each step, the agent receives information about the current state of the dialogue and generates a response.
3. The user provides feedback in the form of a reward, reflecting the quality of the generated response in the context of the current dialogue. The reward function determines how the reward for each action of the agent is calculated, and can take into account various aspects such as the relevance of the response, its grammatical correctness, user satisfaction, etc.
4. The agent updates its strategy in such a way as to maximize the value function – the total expected reward throughout the dialogue.

The advantage of reinforcement learning-based chatbots is their ability to adapt to different user interaction scenarios and optimize their behavior to achieve desired results. In addition, such agents can learn based on implicit feedback from the user, which avoids the need for explicit data labeling.

Reinforcement learning based on human feedback (RLHF (figure 9)) is usually performed iteratively [31, p. 92]:

1. The reward model is initially trained on a dataset of human judgments.
2. The policy model is trained to maximize reward using the current reward model.
3. Humans then evaluate the outputs of the policy model to create a new dataset.
4. This new dataset is used to update the reward model, making it more accurate.
5. The policy model is again fine-tuned using the updated reward model.

Generative Adversarial Networks (GANs) are one of the most innovative approaches to dialogue management in chatbots. GANs consist of two neural networks – a generator and a discriminator, which compete with each other in the learning process. The goal of the generator is to create responses that are so realistic that the discriminator cannot distinguish them from real human responses. The goal of the discriminator is to learn to distinguish between generated responses and real responses.

Chou and Hsueh [33] proposed a method for creating a chatbot using a model that generates sequential sentences based on a generative adversarial network. The model architecture contains a generator

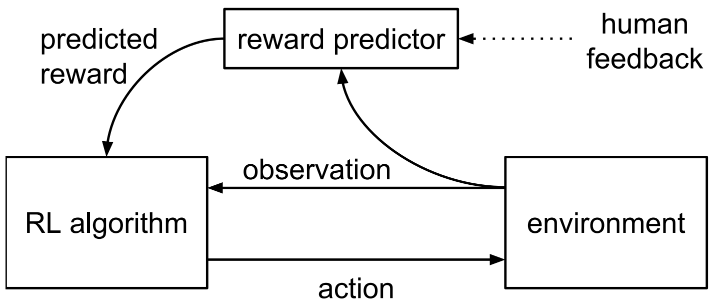


**Figure 9:** Reinforcement learning based on human feedback [32, p. 2].

that generates diverse sentences and a discriminator that evaluates sentences. The generator combines an attention model that responds to tracking conversation states with a Seq2Seq model using LSTM to obtain sentence information. For the discriminator, two types of rewards are calculated to assign low rewards for repetitive sentences and high rewards for diverse sentences. Under this approach, the environment is the model itself.

The model proposed by Tran et al. [34] combines reinforcement learning and generative adversarial networks to generate both accurate and human-like responses (figure 10).

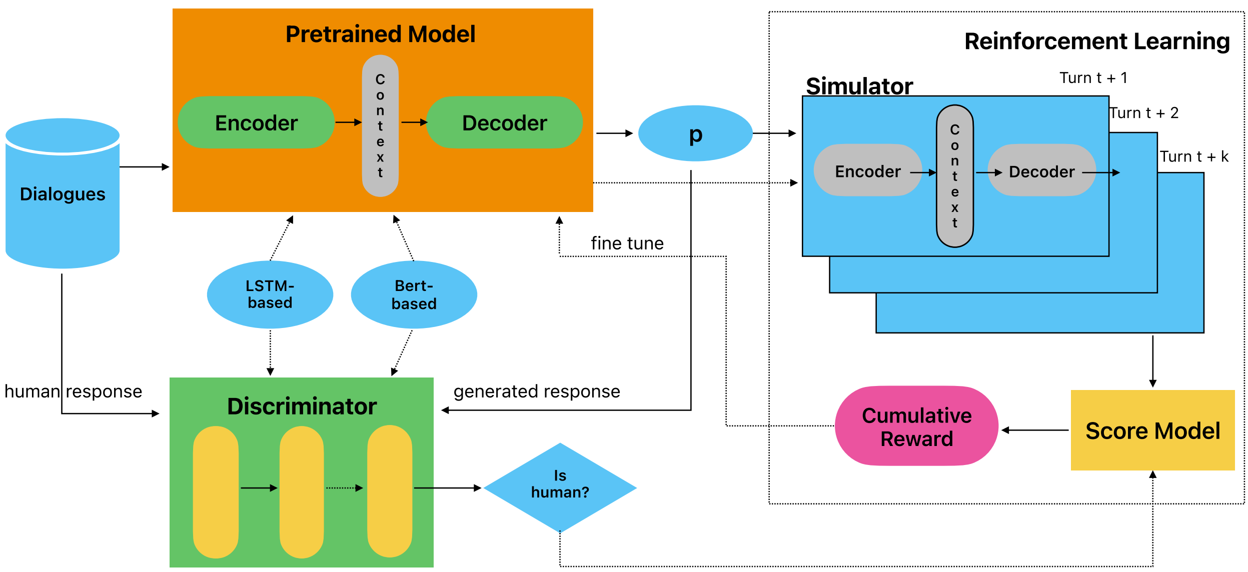


**Figure 10:** Hybrid model combining RLHF and GAN for chatbot training (based on [34, p. 75962]).

## 3.3. Transfer learning

Transfer learning is an approach to machine learning that allows using the knowledge gained by the model when solving one task to improve its efficiency when solving another, similar task [31, p. 81].

The idea of transfer learning is to first train the model on a large amount of data to perform a general task (for example, predicting the next word in a sentence), and then fine-tune it on a smaller amount of data to perform a specific task (for example, generating chatbot responses in a certain subject area) [35].

Adapting models to a specific subject area (domain) is an important area of transfer learning that allows improving the quality of chatbots in specific areas of application. Ilievski et al. [36, p. 4116] distinguish between two cases when two domains overlap and when one domain is an extension of the other, and point out the advantages of using transfer learning for training specialized chatbots:

- training chatbots on a smaller amount of data: in data-constrained environments, models trained using transfer learning achieve better training and testing results than models trained independently;
- better chatbot performance: the use of transfer learning has a significant positive impact on performance, even when all data from the target domain is available.

Fine-tuning existing large language models instead of creating them from scratch is often a more practical and effective approach for the following reasons [31, pp. 80-82]:

- resource efficiency – fine-tuning uses pre-trained models, allowing for high performance without significant investments in computational resources;
- data efficiency – fine-tuning allows leveraging the large amount of data on which the models were pre-trained, requiring only a smaller, task-specific dataset for adaptation;
- transfer learning – pre-trained models have a general understanding of language, context and certain domain knowledge that is transferred to a specific task during fine-tuning;
- high performance – fine-tuning allows using state-of-the-art architectures that have been carefully optimized and tested by industry experts;
- lower entry barrier – fine-tuning is more accessible to organizations and individuals who do not have sufficient infrastructure to train models from scratch;
- continuous learning – pre-trained models can be continuously updated and tuned for different tasks, making them versatile and adaptable;
- wide applicability – one pre-trained model can be fine-tuned for many domains and tasks.

The fine-tuning process includes the following steps (figure 11) [31, pp. 82-84]:

1. Loading the parameters of the pre-trained model.
2. Preparing a task-specific dataset.
3. Extracting features using the pre-trained layers of the model.
4. Adjusting the model parameters through backpropagation and gradient descent with a lower learning rate.
5. Updating gradients, applying regularization to prevent overfitting.
6. Choosing a fine-tuning strategy: full or partial tuning of the model.
7. Evaluating performance and optimizing model hyperparameters.

During fine-tuning of a neural network, changes occur at the level of architecture and individual neurons [31, pp. 85-88]:

- adjusting weights and changing activation function outputs;
- tuning the upper layers of the model and freezing the lower layers;
- adjusting the feature space to adapt to the new task;
- replacing the last layer to match the new task;
- updating batch normalization parameters.

Fine-tuning allows efficiently adapting pre-trained models to specific tasks, preserving their "intuition" and optimizing for the specifics of the new task.

### 3.4. Evaluation of chatbot training effectiveness

Taulli [37] considers a number of performance metrics for large language models (table 3).

Human evaluation is a group of methods that involve direct assessment of the quality of chatbots' work by real users or experts. These methods allow obtaining a more holistic and comprehensive assessment of the model's effectiveness, taking into account aspects such as relevance, coherence, naturalness, and usefulness of the generated responses.

For human evaluation of the quality of a fine-tuned model, the following steps should be performed:

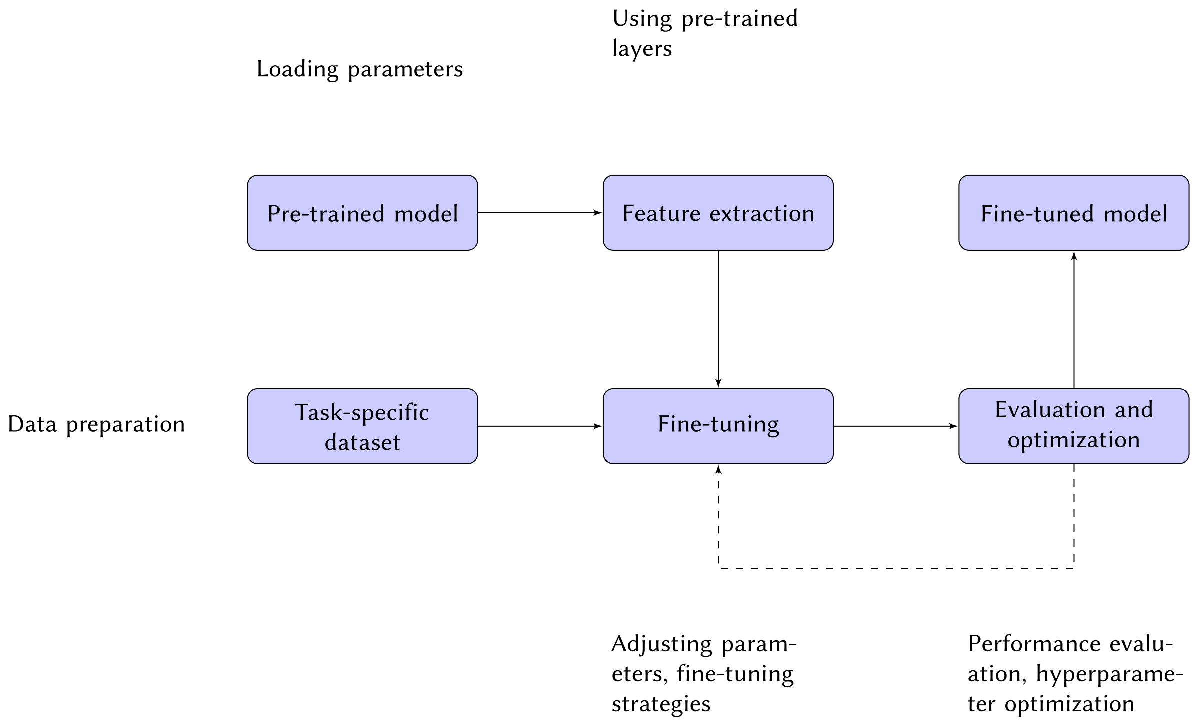


**Figure 11:** Scheme of the fine-tuning process.

**Table 3**
Metrics for evaluating text generation models (based on [37, pp. 36-37]).

| Metric | Description |
|---|---|
| BERTScore | evaluates by comparing the generated text with the reference text using BERT embeddings |
| Perplexity | evaluates how well the probability distribution predicted by the model matches the actual data distribution |
| BLEU | computes n-gram precision scores for similarity between the generated and reference text |
| ROUGE | computes n-gram overlap between the generated and reference text |

1. Use the fine-tuned model to generate new texts.
2. Engage humans (experts in the relevant field or regular users) to evaluate the quality of the generated texts.
3. Develop an evaluation system that takes into account aspects such as grammar, coherence, relevance, style, and overall quality.
4. Collect feedback and ratings from multiple people and average the results to obtain an overall assessment of the model's performance.

# 4. Experiments

## 4.1. Creating datasets for training chatbots

For conducting experiments on training chatbots, two datasets were created containing texts of scientific publications in the field of information technology.

*CEUR Workshop Proceedings* is a Diamond OA publication that publishes materials from scientific conferences and seminars covering a wide range of research in computer science and engineering. The following steps were performed to form the dataset from its publications:

- downloading the website https://ceur-ws.org/ using the `wget` utility;

- extracting texts from 68791 PDF files corresponding to volumes 1-3583 for 1995-2024;
- creating a single text file `ceur-ws.txt` with a size of 1917 MB (2009797694 bytes).

The relative distribution of publications by languages of articles according to Scopus: English – 94.807%, Russian – 1.368%, German – 1.101%, Spanish – 0.772%, Portuguese – 0.691%, Turkish – 0.596%, French – 0.340%, Ukrainian – 0.160%, Italian – 0.131%, Czech – 0.015%, other languages together – 0.019%.

The next dataset was created from publications of the journal *Information Technologies and Learning Tools* that publishes articles on theoretical and applied aspects of the use of information and communication technologies in education. The procedure for creating this dataset included:

- downloading the journal's website https://journal.iitta.gov.ua/index.php/itlt using the `wget` utility;
- extracting texts from 1732 PDF files corresponding to volumes 1-100 of the journal for 2006-2024;
- forming a text file `journal.txt` with a volume of 107 MB (112245281 bytes).

The specificity of this dataset is the narrower subject specialization of texts, which focus on issues of digital transformation of education, development and use of electronic educational resources, technologies of distance and blended learning, etc.

The approximate relative distribution of publications by languages of articles according to Web of Science: Ukrainian – 52.54%, Russian – 26.73%, English – 20.73%.

The formed datasets differ in volume, time coverage, thematic focus of texts and distribution of languages of writing texts, which allows using them for comparative analysis of the effectiveness of chatbot models trained on heterogeneous text corpora.

### 4.2. Selection of models for fine-tuning

Considering the prevalence of models of the GPT (Generative Pre-trained Transformer) family, the possibilities of access to both their modern versions (GPT 3.5, 4.0) and alternative models (Gemini 1.0, Claude 3) were studied:

- OpenAI provides a programmatic and visual [38] interface for fine-tuning models gpt-3.5-turbo-0125, gpt-3.5-turbo-1106, gpt-3.5-turbo-0613, babbage-002, davinci-002, and gpt-4-0613. None of these models are free for fine-tuning – as of the 1st quarter of 2024, the cheapest babbage-002 requires a payment of 0.40$ per 1 million tokens, while more advanced models are 20 or more times more expensive. In addition to the payment for fine-tuning, there is a payment for usage – for example, the cost of input and output messages for the davinci-002 model is 12.00$ per 1 million tokens.
- Google also provides a programmatic and visual [39] interface for fine-tuning models gemini-1.0-pro, gemini-1.5-pro, gemini-1.5-flash, gemini-1.0-pro, but the latter is not available in Ukraine [40].
- Anthropic, as of the 1st quarter of 2024, does not provide the possibility of fine-tuning models of the Claude family (except for selected groups of experts) [41].

Considering the unpredictability of the budget for fine-tuning and using fine-tuned models, instead of modern versions of GPT, the historical GPT-2 model [42] was chosen, trained on an internal dataset of OpenAI – WebText (40 GB). The feasibility of using this model as a basis for a chatbot is shown in the work [43].

Hugging Face provides access to all basic versions of the GPT-2 model via the link https://huggingface.co/openai-community: GPT2, GPT2-Medium, GPT2-Large, and GPT2-XL, trained on English-language texts. Considering the comparability of the dataset from *CEUR Workshop Proceedings* publications (2 GB) and the WebText dataset (40 GB), as well as the predominant language (English), the choice of the GPT2-XL model was reasonable for fine-tuning.

**Table 4**

Ukrainian-language GPT-2 models on Hugging Face (as of 11.06.2024).

| Name | Version | Dataset |
|---|---|---|
| Tereveni-AI/gpt2-124M-uk-fiction | GPT2 | Model was trained on corpus of 4040 fiction books, 2.77 GiB in total |
| kyryl0s/gpt2-uk-xxs | GPT2 | GPT2 being trained on Ukrainian news. |
| benjamin/gpt2-large-wechsel-ukrainian | GPT2-Large | gpt2-large transferred to Ukrainian using the method from the NAACL2022 paper WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. |
| kyryl0s/gpt2-uk-zno-edition | GPT2? | GPT2 trained to generate ZNO (Ukrainian exam SAT type of thing) essays |
| malteos/gpt2-uk | GPT2 | A generative language model for the Ukrainian language follows the GPT-2 architecture (124M parameters). Training data: OSCAR, Wikimedia dumps |
| nvidia/GPT-2B-001 | GPT2-XL | The model was trained on 1.1T tokens obtained from publicly available data sources. The dataset comprises 53 languages and code. |

Hugging Face contains a large (more than 11 thousand) number of GPT-2 models fine-tuned (https://huggingface.co/models?sort=trending&search=gpt2), among which less than 10 were fine-tuned on Ukrainian-language texts (table 4).

Considering the small volume of the dataset from publications of the journal *Information Technologies and Learning Tools* and the predominance of the Ukrainian language in it, it is reasonable to choose for fine-tuning a model that has already been fine-tuned on a Ukrainian-language set of texts. Among the models presented in table 4, malteos/gpt2-uk was chosen for fine-tuning due to its small size (124M parameters) and non-specificity of the OSCAR and Wikimedia datasets compared to the datasets used for fine-tuning other models – fiction texts, news, essays for ZNO.

### 4.3. The process of fine-tuning the chatbot model based on GPT2-uk

For fine-tuning the chatbot model based on GPT2-uk on a Ukrainian-language text corpus, the following Python code using the `transformers` library was implemented in the Google Colab environment:

```
from transformers import TextDataset
from transformers import DataCollatorForLanguageModeling
from transformers import GPT2Tokenizer, GPT2LMHeadModel
from transformers import Trainer, TrainingArguments
```

First, the necessary classes from the `transformers` library are imported for working with text data, tokenizer and GPT2 model, as well as for configuring the training process.

Next, auxiliary functions were defined:

```
def load_dataset(file_path, tokenizer, block_size = 128):
    dataset = TextDataset(
        tokenizer = tokenizer,
        file_path = file_path,
        block_size = block_size,
    )
    return dataset

def load_data_collator(tokenizer, mlm = False):
    data_collator = DataCollatorForLanguageModeling(
```

```
        tokenizer=tokenizer,
        mlm=mlm,
    )
    return data_collator
```

The `load_dataset` function loads a text dataset, performing its tokenization with a given block size `block_size`. The `load_data_collator` function creates an object that will prepare data batches for training the model (with the possibility of masked language modeling).

The main function `train` performs the direct fine-tuning of the model:

```
def train(train_file_path,model_name,
          output_dir,
          overwrite_output_dir,
          per_device_train_batch_size,
          num_train_epochs,
          save_steps, resume_from_checkpoint):
  tokenizer = GPT2Tokenizer.from_pretrained(model_name)
  train_dataset = load_dataset(train_file_path, tokenizer)
  data_collator = load_data_collator(tokenizer)

  tokenizer.save_pretrained(output_dir)

  model = GPT2LMHeadModel.from_pretrained(model_name)

  model.save_pretrained(output_dir)

  training_args = TrainingArguments(
          output_dir=output_dir,
          overwrite_output_dir=overwrite_output_dir,
          per_device_train_batch_size=per_device_train_batch_size,
          num_train_epochs=num_train_epochs,
      )

  trainer = Trainer(
          model=model,
          args=training_args,
          data_collator=data_collator,
          train_dataset=train_dataset,
  )

  trainer.train(resume_from_checkpoint=resume_from_checkpoint)
  trainer.save_model()
```

This function performs the following steps:

1. Loading and saving the tokenizer and model initialized from the specified `model_name` point.
2. Preparing the training dataset and `data_collator` object.
3. Defining training parameters `training_args`: directory for saving the model, batch size, number of iterations.
4. Creating a `trainer` object with the model, training parameters, and training data.
5. Launching model training using the `train()` method with the ability to continue from a checkpoint.

6. Saving the fine-tuned model.

The fine-tuning parameters are set by the following variables:

```
#original model and checkpoints
train_directory =
   "/content/drive/MyDrive/SemerikovProject/q_and_a"
train_file_path =
   "/content/drive/MyDrive/SemerikovProject/journal.txt"
model_name = train_directory
#fine-tuned model and checkpoints
output_dir =
   "/content/drive/MyDrive/SemerikovProject/custom_full_text"
overwrite_output_dir = False
per_device_train_batch_size = 8
num_train_epochs = 50
save_steps = 50000
```

The `train_file_path` variable specifies the path to the text file with the fine-tuning corpus, `model_name` – the directory with the files of the initial model (or the checkpoint for continuing fine-tuning), `output_dir` – the output directory for saving the fine-tuned model files and checkpoints.

The initial model `malteos/gpt2-uk` is loaded using the methods of the `transformers` library from the Hugging Face hub:

```
from transformers import AutoTokenizer, AutoModelForCausalLM

tokenizer = AutoTokenizer.from_pretrained("malteos/gpt2-uk")
model = AutoModelForCausalLM.from_pretrained("malteos/gpt2-uk")
tokenizer.save_pretrained(train_directory)
model.save_pretrained(train_directory)
```

The loaded model files are saved to the specified `train_directory`.

The fine-tuning process is launched by calling the `train` function with the specified parameters:

```
train(train_file_path=train_file_path,
        model_name=model_name,
        output_dir=output_dir,
        overwrite_output_dir=overwrite_output_dir,
        per_device_train_batch_size=per_device_train_batch_size,
        num_train_epochs=num_train_epochs,
        save_steps=save_steps,
        # False for the first time,
        # True - to continue after resume
        resume_from_checkpoint=False)
```

The training time can be reduced by using a GPU on a separate server or local computer. The latter required installing the CUDA 12.5 driver set to support the GeForce RTX 3080 (10 Gb) video card and a specialized version of PyTorch, as well as updating the accelerate and transformers libraries.

For the final experiment, the total training time was 37 hours (the average duration of one training iteration was 43 minutes).

To test the model fine-tuned on the dataset, a user interface was created using the `gradio` library (figure 12, https://huggingface.co/spaces/POMAHSLS/ITLT_Journal):

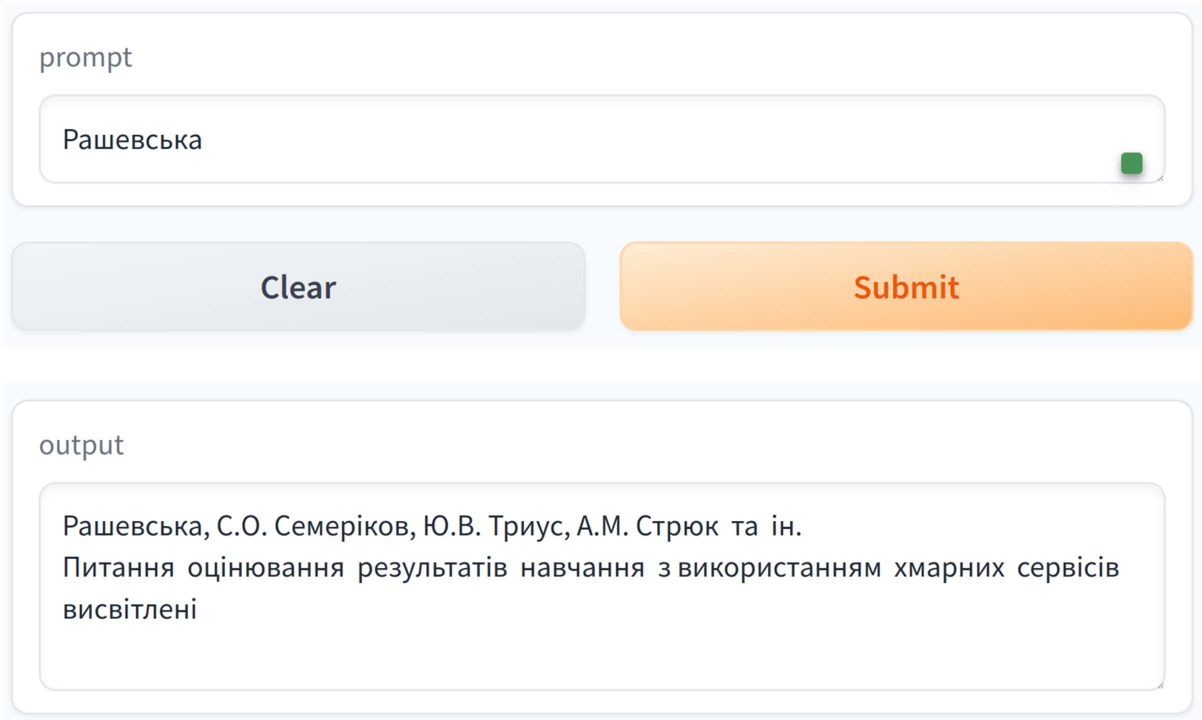


**Figure 12:** User interface prototype of the chatbot for the GPT-2 model fine-tuned on the dataset from the publications of the journal "Information Technologies and Learning Tools".

```python
from transformers import pipeline
import gradio as gr

model = pipeline("text-generation",
  model="/content/drive/MyDrive/SemerikovProject/lastversion")

def predict(prompt):
    completion = model(prompt, max_length=50)[0]["generated_text"]
    return completion

gr.Interface(fn=predict, inputs="text", outputs="text").launch()
```

## 5. Conclusions

1. Bibliometric analysis of 549 sources from the Scopus database on the problem of chatbot training provided an opportunity to:
   a) identify the lower chronological boundary (2018), starting from which there is a steady increase in the number of publications on chatbot training;
   b) suggest that the reason for a significant increase in the number of works in 2023 (by 55 compared to 2022) is the public availability of ChatGPT and the associated surge of interest from the scientific community in the topic of large language models and chatbots: to confirm or refute this assumption, one can analyze the context of keywords in 2023 publications for direct mentions of ChatGPT or similar systems, and also consider the dynamics of citations of works dedicated to ChatGPT in other works during this period;
   c) group into 4 clusters the author and index keywords of sources on chatbot training: 1) natural language processing; 2) application of natural language processing technologies in society;

3) application of machine learning for natural language processing; 4) chatbots in education and service sector.

2. To determine the most significant key concepts of the research in each cluster, the keywords that have the largest and the next largest value were identified by the following indicators: number of links, total link strength, number of documents with given keyword and averages: publication year, number of citations, normalized number of citations. For the selected keywords, their significance was calculated from 1 to 6:
   a) in the first cluster (natural language processing), the most significant are 7 keywords (41%): computational linguistics (6), natural language understanding (3), performance (2), language model (2), speech processing (1);
   b) in the second cluster (application of natural language processing technologies in society), the most significant are 10 keywords (63%): artificial intelligence (6), large language model (2), chatgpt (2), human (2), training (2), review (2), conversational agent (1), controlled study (1), education (1), male (1);
   c) in the third cluster (application of machine learning for natural language processing), the most significant are 6 keywords (50%): natural language processing (6), natural language processing systems (3), virtual assistant (3), language processing (2), user interfaces (2), diagnosis (2);
   d) in the fourth cluster (chatbots in education and service sector), the most significant are 7 keywords (58%): chatbot (6), learning systems (3), curricula (3), learn+ (2), knowledge based systems (2), personnel training (1), customer service (1).
3. During the analysis of the keyword map, it was found that the third cluster (application of machine learning for natural language processing) has certain intersections with the content of the first (natural language processing) and second (application of natural language processing technologies in society) clusters. This is explained by the fact that machine learning methods, particularly deep learning, are fundamental tools for developing natural language processing systems and their applications. Therefore, some key terms of the third cluster, such as natural language processing, virtual assistants, neural networks, etc., are closely related to the concepts of other clusters. This indicates the close integration of different research areas in the single scientific field of creating intelligent conversational agents. Further analysis of the intersections between clusters may become a promising direction for future scientific research to more clearly delineate subject areas.
4. Analysis of the map of connections of the most significant keywords provided an opportunity to identify the leading research directions:
   a) in natural language processing using computational linguistics – natural language understanding, construction of language models and speech recognition;
   b) in the application of artificial intelligence technologies for natural language processing – controlled use of large language models and chatbots (in particular, ChatGPT) in education;
   c) in the application of machine learning for natural language processing – the use of virtual assistants, natural language user interfaces and other natural language processing systems, in particular, for diagnosis;
   d) in the application of chatbots in education and service sector – the use of chatbots, learning systems and knowledge management systems for enhanced and adaptive learning.

   The identified directions can be useful for organizations in developing strategies for using artificial intelligence and integrating conversational agents into production activities.
5. The results of the conducted bibliometric analysis can be applied:
   - in education:
     a) the identified priority areas of research in the field of chatbot training can be used to form the topics of courses, training programs for specialists in natural language processing and development of artificial intelligence systems;

   b) the identified key concepts can serve as a basis for the development of educational materials designed to highlight the most important concepts and technologies for creating conversational agents;
   c) an overview of the applications of chatbots in the educational process outlines promising ways of their integration into adaptive and personalized learning systems;
 - in scientific research:
   a) the map of connections of key concepts can serve as a basis for forming scientific hypotheses and constructing conceptual models during research design;
   b) the results of the analysis open up prospects for further bibliometric studies of dynamics and relationships in the subject field;
   c) the identified priority areas determine the relevant vectors of future scientific research in the field of artificial intelligence and its use in scientific research in the field of social sciences.

6. Supervised learning is one of the main approaches that involves training the model on labeled "query-response" pairs. For this, architectures based on recurrent neural networks (for example, Seq2Seq with LSTM) and transformers (for example, GPT) are used. These models are capable of generating contextually relevant responses, but require large volumes of high-quality labeled data.
7. Reinforcement learning allows the model to learn through interaction with the environment (user) and receive feedback in the form of rewards. This approach is implemented using generative adversarial networks (GAN) and an iterative process of reinforcement learning based on human feedback (RLHF).
8. Transfer learning consists of using the knowledge gained by the model when solving one task to improve its efficiency when solving another, similar task. The most common approach is fine-tuning a pre-trained model on a specific dataset to adapt to a specific subject area.
9. To evaluate the effectiveness of chatbot training, both automatic metrics (BERTScore, perplexity, BLEU, ROUGE) and human evaluation methods are used, which allow taking into account the relevance, coherence and naturalness of the generated responses.
10. Two datasets for training chatbots were formed: a dataset from "CEUR Workshop Proceedings" publications (a wide range of research in computer science in English) and a dataset from publications of the journal "Information Technologies and Learning Tools" (a subject-oriented corpus mainly in Ukrainian). The created datasets differ in volume, time coverage, thematic focus and distribution of languages of writing texts, which provides conditions for a comparative analysis of the effectiveness of chatbot models trained on heterogeneous text corpora.
11. The choice of models for fine-tuning was substantiated: the basic multilingual GPT2-XL model for the first dataset and the gpt2-uk model previously fine-tuned on Ukrainian texts for the second. The selection of models was based on comparing the characteristics of text datasets (volume, language composition) and available models of the GPT family, taking into account their size and previous training experience.
12. The process of fine-tuning the selected models on the formed text corpora using the capabilities of the `transformers` library from Hugging Face was implemented. The developed program code allows loading pre-trained models, fine-tuning them on text data provided by the user, and saving fine-tuned models for further use.
13. The operation of the fine-tuned models for generating chatbot responses to thematically related user queries was tested. For the convenience of user interaction with the chatbot, a graphical interface was developed using the `gradio` library.

The obtained results can be used for further research in the direction of creating effective and specialized chatbots using modern approaches to training large language models.

**Acknowledgments:** The authors would like to thank the developers of the Hugging Face transformers library and the providers of the GPT-2 and gpt2-uk models for making their work publicly available. We also express our gratitude to the publishers of the "CEUR Workshop Proceedings" and "Information Technologies and Learning Tools" for maintaining open access to their valuable scientific content.

**Declaration on Generative AI:** During the preparation of this work, the authors used Claude 3 Opus in order to: Text Translation, Abstract drafting, Formatting assistance. After using this tool, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

## References

[1] G. M. Franklin, Google's new AI Chatbot produces fake health-related evidence-then self-corrects, PLOS Digital Health 3 (2024) 1–4. doi:10.1371/journal.pdig.0000619.

[2] A. G. Usigan, M. I. Salomeo, G. J. L. J. Zafe, C. J. Centeno, A. A. R. C. Sison, A. G. Bitancor, Implementation of an Undergraduate Admission Chatbot Using Microsoft Azure's Question Answering and Bot Framework, in: Proceedings of the 2022 5th Artificial Intelligence and Cloud Computing Conference, AICCC '22, Association for Computing Machinery, New York, NY, USA, 2023, p. 240–245. doi:10.1145/3582099.3582135.

[3] R. Thamilselvan, P. Natesan, A. Manimaran, S. E. Naveenkumar, J. K. Shanthosh, S. Vigneshwaran, Designing A Llama 2-Powered Chatbot for Enhanced College Website Support, in: 2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2024, pp. 1–6. doi:10.1109/ICCCNT61001.2024.10725472.

[4] A. D. Workman, V. K. Rathi, D. K. Lerner, J. N. Palmer, N. D. Adappa, N. A. Cohen, Utility of a LangChain and OpenAI GPT-powered chatbot based on the international consensus statement on allergy and rhinology: Rhinosinusitis, International Forum of Allergy & Rhinology 14 (2024) 1101–1109. doi:10.1002/alr.23310.

[5] OpenAI, Introducing ChatGPT, 2022. URL: https://openai.com/blog/chatgpt.

[6] S. V. Symonenko, N. V. Zaitseva, V. V. Osadchyi, K. P. Osadcha, V. S. Kruglyk, S. O. Sysoieva, Application of chatbots for enhancing communication skills of IT specialists, Journal of Physics: Conference Series 2871 (2024) 012026. doi:10.1088/1742-6596/2871/1/012026.

[7] A. V. Riabko, T. A. Vakaliuk, O. V. Zaika, R. P. Kukharchuk, V. V. Kontsedailo, Chatbot algorithm for solving physics problems, CEUR Workshop Proceedings 3553 (2023) 75–92.

[8] I. Mintii, S. Semerikov, Optimizing Teacher Training and Retraining for the Age of AI-Powered Personalized Learning: A Bibliometric Analysis, in: E. Faure, Y. Tryus, T. Vartiainen, O. Danchenko, M. Bondarenko, C. Bazilo, G. Zaspa (Eds.), Information Technology for Education, Science, and Technics, volume 222 of *Lecture Notes on Data Engineering and Communications Technologies*, Springer Nature Switzerland, Cham, 2024, pp. 339–357. doi:10.1007/978-3-031-71804-5_23.

[9] R. Liashenko, S. Semerikov, The Determination and Visualisation of Key Concepts Related to the Training of Chatbots, in: E. Faure, Y. Tryus, T. Vartiainen, O. Danchenko, M. Bondarenko, C. Bazilo, G. Zaspa (Eds.), Information Technology for Education, Science, and Technics, volume 222 of *Lecture Notes on Data Engineering and Communications Technologies*, Springer Nature Switzerland, Cham, 2024, pp. 111–126. doi:10.1007/978-3-031-71804-5_8.

[10] R. Liashenko, S. Semerikov, Bibliometric analysis of chatbot training research: Key concepts and trends, Information Technologies and Learning Tools 101 (2024) 181–199. doi:10.33407/itlt.v101i3.5622.

[11] DeepLearning.AI, Search | The Batch | AI News & Insights, 2023. URL: https://www.deeplearning.ai/search/.

[12] Big Bot Makes Small Talk: A research summary of Facebook's Generative BST chatbot, 2020. URL: https://www.deeplearning.ai/the-batch/big-bot-makes-small-talk/.

[13] Bot Comic: How Google's Meena chatbot developed a sense of humor, 2020. URL: https://www.deeplearning.ai/the-batch/bot-comic/.

[14] Chatbots for Productivity: Microsoft extends Copilot to 365 and Windows, 2023. URL: https://www.deeplearning.ai/the-batch/microsoft-extends-copilot-365-windows/.

[15] China Chases Chatbots: Chinese tech companies race to cash in on ChatGPT fever, 2023. URL: https://www.deeplearning.ai/the-batch/chinese-tech-companies-race-to-cash-in-on-chatgpt-fever/.

[16] Search War! Google and Microsoft both announce AI-Powered search, 2023. URL: https://www.deeplearning.ai/the-batch/google-and-microsoft-both-announce-ai-powered-search/.

[17] Chatbots Disagree on Covid-19: Medical chatbots offered conflicting Covid advice, 2020. URL: https://www.deeplearning.ai/the-batch/chatbots-disagree-on-covid-19/.

[18] Language Models, Extended: Large language models grew more reliable and less biased in 2022, 2022. URL: https://www.deeplearning.ai/the-batch/language-models-grew-more-reliable-and-less-biased-in-2022/.

[19] Cost Containment for Generative AI: Microsoft's quest to reduce the size and cost of language models, 2023. URL: https://www.deeplearning.ai/the-batch/microsofts-quest-to-reduce-the-size-and-cost-of-language-models/.

[20] What We Know — and Don't Know — About Foundation Models: A new Stanford index to assess the transparency of leading AI models, 2023. URL: https://www.deeplearning.ai/the-batch/a-new-stanford-index-to-assess-the-transparency-of-leading-ai-models/.

[21] Elsevier B.V., Scopus - Document search | Signed in, 2023. URL: https://www.scopus.com/search/form.uri?display=basic#basic.

[22] N. J. Van Eck, L. Waltman, VOSviewer Manual, Universiteit Leiden, 2023. URL: https://www.vosviewer.com/documentation/Manual_VOSviewer_1.6.20.pdf.

[23] Centre for Science and Technology Studies, Leiden University, The Netherlands, VOSviewer - Visualizing scientific landscapes, 2023. URL: https://www.vosviewer.com/.

[24] S. P. Uprety, S. R. Jeong, The Impact of Semi-Supervised Learning on the Performance of Intelligent Chatbot System, Computers, Materials & Continua 71 (2022) 3937–3952. doi:10.32604/cmc.2022.023127.

[25] S. Patil, V. Mudaliar, P. Kamat, LSTM based Ensemble Network to enhance the learning of Long-term Dependencies in Chatbot, International Journal of Automation and Smart Technology 12 (2022) 2286–2286. doi:10.5875/ausmt.v12i1.2286.

[26] Z. Ji, A Multi-modal Seq2seq Chatbot Framework, in: Z. Qian, M. Jabbar, X. Li (Eds.), Proceeding of 2021 International Conference on Wireless Communications, Networking and Applications, Springer Nature, Singapore, 2022, pp. 225–233. doi:10.1007/978-981-19-2456-9_24.

[27] P. Anki, A. Bustamam, H. S. Al-Ash, D. Sarwinda, Intelligent Chatbot Adapted from Question and Answer System Using RNN-LSTM Model, Journal of Physics: Conference Series 1844 (2021) 012001. doi:10.1088/1742-6596/1844/1/012001.

[28] D. Dharrao, S. Gite, TherapyBot: a chatbot for mental well-being using transformers, International Journal of Advances in Applied Sciences 13 (2024) 1–12. doi:10.11591/ijaas.v13.i1.pp1-12.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention Is All You Need, 2023. arXiv:1706.03762.

[30] M. Lee, A Mathematical Investigation of Hallucination and Creativity in GPT Models, Mathematics 11 (2023) 2320. doi:10.3390/math11102320.

[31] A. Kansal, Finetuning: The Theory, in: Building Generative AI-Powered Apps: A Hands-on Guide for Developers, Apress, Berkeley, CA, 2024, pp. 77–100. doi:10.1007/979-8-8688-0205-8_5.

[32] P. Christiano, J. Leike, T. B. Brown, M. Martic, S. Legg, D. Amodei, Deep reinforcement learning from human preferences, 2023. arXiv:1706.03741.

[33] T.-L. Chou, Y.-L. Hsueh, A Task-oriented Chatbot Based on LSTM and Reinforcement Learning, in: Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval, NLPIR '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 87–91. doi:10.1145/3342827.3342844.

[34] Q.-D. L. Tran, A.-C. Le, V.-N. Huynh, Enhancing Conversational Model With Deep Reinforcement Learning and Adversarial Learning, IEEE Access 11 (2023) 75955–75970. doi:10.1109/ACCESS.2023.3297652.

[35] R. Liashenko, S. Semerikov, Training Specialised Chatbots on Ukrainian Scientific Text Corpora Using Transfer Learning, in: 2024 IEEE 18th International Conference on Computer Science and

Information Technologies (CSIT), IEEE, 2025 (in press).
[36] V. Ilievski, C. Musat, A. Hossman, M. Baeriswyl, Goal-oriented chatbot dialog management bootstrapping with transfer learning, in: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18, International Joint Conferences on Artificial Intelligence Organization, 2018, pp. 4115–4121. doi:10.24963/ijcai.2018/572.
[37] T. Taulli, AI-Assisted Programming: Better Planning, Coding, Testing, and Deployment, O'Reilly Media, Inc., Sebastopol, CA, 2024. URL: https://www.oreilly.com/library/view/ai-assisted-programming/9781098164553/.
[38] OpenAI, Fine-tuning - openai api, 2024. URL: https://platform.openai.com/finetune.
[39] Google, Google ai studio, 2024. URL: https://aistudio.google.com/.
[40] Google, Available regions for Google AI Studio and Gemini API, 2024. URL: https://ai.google.dev/gemini-api/docs/available-regions.
[41] S. Elaprolu, Introducing the AWS Generative AI Innovation Center's Custom Model Program for Anthropic Claude, 2023. URL: https://aws.amazon.com/blogs/machine-learning/introducing-the-aws-generative-ai-innovation-centers-custom-model-program-for-anthropic-claude/.
[42] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language Models are Unsupervised Multitask Learners, 2019. URL: https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.
[43] A. W. Ohlheiser, K. Hao, An AI is training counselors to deal with teens in crisis, 2021. URL: https://www.technologyreview.com/2021/02/26/1020010/trevor-project-ai-suicide-hotline-training/.

# A. Results of applying generative chatbots to determine cluster names

The query was created in English on November 30, 2023. The query contained information about the number of clusters and the keywords related to each cluster. The chatbot was expected to make suggestions regarding the names and contents of the clusters. The query had the following form:

"There are 4 clusters of keywords related to chatbot training research:

Cluster 1: classification (of information), training data, speech processing, performance, semantics, dialogue systems, natural language understanding, computational linguistics, text processing, long short-term memory, state of the art, language model, question answering, information retrieval, embeddings, reinforcement learning, response generation

Cluster 2: artificial intelligence, conversational agent, human, adult, training, male, medical education, female, controlled study, health care, mental health, chatgpt, education, covid-19, review, large language model

Cluster 3: natural language processing, natural language processing systems, machine learning, deep learning, natural languages, language processing, learning algorithms, virtual assistant, user interfaces, convolutional neural network, query processing, diagnosis

Cluster 4: chatbot, learning systems, students, e-learning, learn+, sales, human computer interaction, knowledge based systems, personnel training, curricula, customer service, engineering education

Please suggest cluster names and brief descriptions for each of the 4 clusters based on the keywords."

### ChatGPT (GPT 3.5)

1. *Advanced Language Processing and AI Applications*: This cluster focuses on advanced language processing techniques and various AI applications related to text and speech. It encompasses topics like natural language understanding, computational linguistics, information retrieval, dialogue systems, and the utilization of cutting-edge technologies such as long short-term memory (LSTM), embeddings, and reinforcement learning in language models and question answering systems.

2. *AI in Healthcare and Social Context*: This cluster centers around the applications of artificial intelligence in healthcare, mental health, and education with a focus on human interaction. It explores the use of conversational agents, particularly in medical education, mental health care, and addressing issues like COVID-19. Additionally, it touches on the gender aspect in AI studies by mentioning controlled studies with adults and gender variations.
3. *Core Technologies in Natural Language Processing and Machine Learning*: This cluster revolves around foundational technologies crucial in natural language processing and machine learning. It includes topics such as deep learning, machine learning algorithms, language processing, neural networks (like convolutional neural networks), query processing, and the development of virtual assistants and user interfaces.
4. *Educational and Service-oriented AI Applications*: This cluster focuses on AI applications in education, learning systems, and service-oriented domains. It highlights topics like e-learning, knowledge-based systems, human-computer interaction, customer service, and sales. It also touches upon the use of AI in student learning, engineering education, and personnel training, emphasizing knowledge acquisition and application in various educational and service-based contexts.

## Bing, precise mode (GPT 4.0)

1. *Language Processing and Information Retrieval*: This cluster focuses on the technical aspects of natural language understanding and computational linguistics. It includes topics such as speech processing, semantics, dialogue systems, text processing, language modeling, question answering, and reinforcement learning for response generation.
2. *AI in Healthcare and Education*: This cluster revolves around the application of artificial intelligence, particularly conversational agents like ChatGPT, in fields such as medical education, healthcare, and mental health. It also touches on demographic factors (male, female, adult) and current issues like COVID-19.
3. *Machine Learning and User Interaction*: This cluster is centered on machine learning, deep learning, and natural language processing systems. It also includes user interfaces and virtual assistants, indicating a focus on how users interact with these AI systems.
4. *AI in Education and Customer Service*: This cluster pertains to the use of AI systems like chatbots in educational settings (e-learning, engineering education, curricula) and customer service. It also includes topics related to human-computer interaction and knowledge-based systems.

## Bing, balanced mode (GPT 4.0)

1. *Natural Language Processing (NLP)*: This cluster includes topics related to the processing of human language by computers. It covers a wide range of subfields such as computational linguistics, text processing, information retrieval, and question answering. The cluster also includes techniques such as embeddings, long short-term memory, and reinforcement learning.:
2. *Artificial Intelligence (AI)*: This cluster includes topics related to the development of intelligent agents that can perform tasks that typically require human intelligence. It covers a wide range of subfields such as conversational agents, large language models, and chatbots. The cluster also includes topics related to health care, education, and COVID-19.:
3. *Machine Learning (ML)*: This cluster includes topics related to the development of algorithms that can learn from data and make predictions or decisions. It covers a wide range of subfields such as deep learning, learning algorithms, and convolutional neural networks. The cluster also includes topics related to virtual assistants, user interfaces, and diagnosis.:
4. *Learning Systems (LS)*: This cluster includes topics related to the development of systems that can learn from data and improve over time. It covers a wide range of subfields such as e-learning, personnel training, and customer service. The cluster also includes topics related to sales, human-computer interaction, and engineering education.:

### Bing, creative mode (GPT 4.0)

1. *Natural Language Understanding and Generation*: This cluster contains keywords related to the tasks and methods of natural language understanding and generation, such as classification, dialogue systems, question answering, and response generation. It also includes keywords that indicate the performance and state of the art of these tasks, such as semantics, long short-term memory, embeddings, and reinforcement learning.
2. *Artificial Intelligence and Health Care*: This cluster contains keywords related to the applications and implications of artificial intelligence in health care, such as conversational agent, medical education, mental health, and covid-19. It also includes keywords that indicate the characteristics and challenges of these applications, such as human, adult, male, female, controlled study, review, and large language model.
3. *Natural Language Processing and Machine Learning*: This cluster contains keywords related to the general concepts and techniques of natural language processing and machine learning, such as natural language processing systems, machine learning, deep learning, learning algorithms, and convolutional neural network. It also includes keywords that indicate the interfaces and domains of these techniques, such as virtual assistant, user interfaces, query processing, and diagnosis.
4. *Chatbot and Education*: This cluster contains keywords related to the design and evaluation of chatbot systems for education, such as chatbot, learning systems, students, e-learning, learn+, and curricula. It also includes keywords that indicate the contexts and outcomes of these systems, such as sales, human computer interaction, knowledge based systems, personnel training, customer service, and engineering education.

### Claude 2

1. *Natural Language Processing Applications*: This cluster focuses on various NLP applications like classification, speech processing, dialogue systems, question answering, etc. Central themes include semantics, computational linguistics, LSTMs, performance, and building practical natural language systems.
2. *Conversational AI and Impacts*: This cluster deals with conversational AI, its use in healthcare/education, and analyzing its impacts. Key themes look at chatbots/assistants, how they interact with and impact humans, with analysis around ethics, proper training, and responsible AI.
3. *Core AI and ML Techniques*: This cluster covers fundamental AI and ML techniques like natural language processing, machine learning, deep learning, algorithms, etc. It focuses on the key technical methods and capabilities that power many AI systems.
4. *AI for Training and Customer Service*: The fourth cluster deals with leveraging AI for use cases like education/training and customer service. Central ideas look at e-learning, knowledge systems, training curriculums, human-computer interaction, and providing automation for repetitive customer service queries.

### Google Bard

1. *Technical NLP*: This cluster focuses on the technical aspects of natural language processing (NLP), including classification, training data, speech processing, semantics, dialogue systems, and language models.
2. *Applications of NLP*: This cluster focuses on the applications of NLP in various domains, including conversational agents, healthcare, education, and customer service.
3. *Machine Learning and Deep Learning for NLP*: This cluster focuses on the machine learning and deep learning techniques used in NLP.
4. *Chatbots and Learning Systems*: This cluster focuses on the development and evaluation of chatbots and other learning systems.