UDC [303.6+303.7]:001.8 **DOI: 10.18413/2518-1092-2026-11-2-0-8**

Chigarev B.N.

MAPPING THE SCHOLARLY LANDSCAPE: A BIBLIOMETRIC ANALYSIS OF AI AND EMERGING TECHNOLOGIES ON ARXIV

Institute of Oil and Gas Problems Russian Academy of Sciences,
3 Gubkina St., Moscow, 119333, Russia

e-mail: bchigarev@ipng.ru

Abstract
Artificial intelligence and emerging technologies are boosting research efficiency and global competitiveness, transforming industries. They also raise important social and ethical governance issues that require evidence-based policy decisions. This article presents a two-stage method for identifying relevant research topics in a proof-of-concept format. The first stage involves analyzing bibliometric records of preprints to identify key topics described by terms taken from abstracts. In the second stage, examples of relevant peer-reviewed publications are identified based on these key terms. This method provides a balance between a broad search for relevant topics and reliable verification of scientific results. The data sources used are metadata from ArXiv preprints on artificial intelligence (cs.AI 126,363 records) and emerging technologies (cs.ET 4,497 records) for 2021–2025. The study used 46,493 multi-word terms found in the annotations of cs.AI bibliometric records. To identify relevant peer-reviewed publications, it is advisable to use artificial intelligence-based search engines such as Semantic Scholar, Elicit, or ScienceOS to search for publications using the terminology identified in the first stage. The study shows that the use of a controlled lexicon allows for the identification of 4–5 groups of interpretable topics in text of abstracts, emphasizing the importance of using terms consisting of 2–4 words to achieve optimal results. Optimizing tasks using computing systems based on physical modeling principles supplemented by artificial intelligence could be a promising area of research.
Keywords: bibliometric analysis; artificial intelligence; promising technologies; preprint metadata; key terms; peer-reviewed publications
Acknowledgements: This work was funded by the Ministry of Science and Higher Education of the Russian Federation, State Assignment no. 125021302095-2
For citation: Chigarev B.N. Mapping the Scholarly Landscape: A Bibliometric Analysis of AI and Emerging Technologies on ArXiv // Research result. Information technologies. – T.11, №2, 2026. – P. 91-108. DOI: 10.18413/2518-1092-2026-11-2-0-8

Чигарев Б.Н.

КАРТИРОВАНИЕ НАУЧНОГО ЛАНДШАФТА: БИБЛИОМЕТРИЧЕСКИЙ АНАЛИЗ ИССЛЕДОВАНИЙ ИИ И ПЕРСПЕКТИВНЫХ ТЕХНОЛОГИЙ НА БАЗЕ ARXIV

Институт проблем нефти и газа РАН,
ул. Губкина, 3, г. Москва, 119333, Россия

e-mail: bchigarev@ipng.ru

Аннотация
Искусственный интеллект и перспективные технологии повышают эффективность научных исследований и глобальную конкурентоспособность, преобразуя отрасли промышленности. Они также поднимают важные социальные и этические вопросы управления, которые требуют принятия политических решений, основанных на фактах. В данной статье представлен двухэтапный метод выявления актуальных тем исследований в формате «доказательства концепции». Первый этап включает анализ библиометрических записей препринтов с целью выявления ключевых тем, описанных терминами, взятыми из аннотаций. На втором этапе на основе этих ключевых терминов выявляются примеры актуальных рецензируемых публикаций. Данный метод обеспечивает баланс между

широким поиском актуальных тем и надежной проверкой научных результатов. В качестве источников данных используются метаданные препринтов ArXiv по искусственному интеллекту (cs.AI 126 363 записи) и перспективные технологиям (cs.ET 4497 записей) за 2021–2025 годы. В исследовании использовались 46 493 многословных термина, найденных в аннотациях библиометрических записей cs.AI. Для выявления соответствующих рецензируемых публикаций рекомендуется использовать поисковые системы на основе искусственного интеллекта, такие как Semantic Scholar, Elicit или ScienceOS, чтобы найти публикации с использованием терминологии, определенной на первом этапе. Исследование показывает, что использование контролируемого лексикона позволяет выделить 4–5 групп интерпретируемых тем в текстах аннотаций, подчеркивая важность использования терминов, состоящих из 2–4 слов, для достижения оптимальных результатов. Оптимизация задач с использованием вычислительные системы на принципах физического моделирования, дополненных искусственным интеллектом, может стать перспективной темой.

Ключевые слова: библиометрический анализ; искусственный интеллект; перспективные технологии; метаданные препринтов; ключевые термины; рецензируемые публикации
Финансирование: Работа выполнена в рамках государственного задания ИПНГ РАН (тема № 125021302095–2).
Для цитирования: Чигарев Б.Н. Картирование научного ландшафта: библиометрический анализ исследований ИИ и перспективных технологий на базе ArXiv // Научный результат. Информационные технологии. – Т.11, №2, 2026. – С. 91-108. DOI: 10.18413/2518-1092-2026-11-2-0-8

INTRODUCTION

Relevance of the topic Artificial Intelligence and Emerging Technologies

Artificial Intelligence (AI) and Emerging Technologies significantly influence scientific progress, improving the efficiency of research and innovation in various disciplines. In economics, they are changing the landscape of industries and global competitiveness [1]. From a scientific point of view, AI serves as a mechanism for formulating and testing hypotheses, accelerating research [2]. The rapid development of technology also raises pressing social, ethical, and governance issues that require sound, evidence-based policies [3].

Importance of Bibliometric Analysis for AI & Emerging Technologies

Bibliometric analysis plays a key role in understanding contemporary challenges and trends in the dynamic field of AI & Emerging Technologies. This method provides a systematic and quantitative basis for evaluating the extensive literature published on the subject, in contrast to traditional qualitative reviews. By conducting a quantitative analysis of publication metadata, including citations, keywords, authors, and institutions, bibliometrics provides an objective picture of publication activity [4]. In addition, analyzing the co-occurrence of keywords allows researchers to identify new “hot topics” and determine less-studied areas or gaps in research that require further investigation [5, 6].

When conducting bibliometric analysis in the field of artificial intelligence and new technologies, Scopus and Web of Science (WoS) are invariably the main and preferred sources of data, while Arxiv is used much less frequently [7]. This is understandable, since Scopus and WoS are commercial, curated databases that mainly index articles from authoritative peer-reviewed journals and conference proceedings. The peer review process is important for ensuring the quality and reliability of academic analysis. Arxiv, on the other hand, serves as a preprint server that allows for the rapid dissemination of research results but does not guarantee their quality, making it less suitable for the final selection of publications. However, in the preliminary stage of analyzing current topics, Arxiv provides a good overview of current research topics and new algorithms, allowing researchers to quickly familiarize themselves with the full texts of new publications. Moreover, industry R&D is often limited to preprints and does not publish articles in peer-reviewed journals.

ArXiv serves as an important source of preprints in the field of computer science, indexed by significant bibliographic databases such as Scopus and Web of Science. They are of considerable value because they provide insight into current research topics, as they are usually made publicly available 1–2 years before the corresponding peer-reviewed articles appear.

The current trajectory of AI R&D shows that private companies operating in the digital economy, such as Google and Facebook, are playing an increasingly prominent role in fundamental research that was previously the preserve of academia. For example, at the 2022 Neural Information Processing Systems (NeurIPS) conference, the largest annual conference on AI/ML, Google (including DeepMind), Microsoft, and Meta accounted for 12.60% (361) of all accepted papers (2,866), more than twice as many as the second most represented institution, Stanford University [8]. The majority of R&D funding for AI originates from the private sector. In 2022, just 100 companies funded 40% of all AI R&D. Large firms with over 250 employees perform 89% of private-sector AI R&D, indicating that industry has the capital and infrastructure to lead the field.

This article proposes a two-stage approach to identifying relevant research topics in a proof-of-concept format. The first stage involves analyzing numerous bibliometric records of preprints to identify widely represented research topics. The phraseology used is identified from the abstracts of these same data. In the second stage, based on the previously identified keywords, relevant implementations are found in peer-reviewed publications. This approach provides an optimal combination of speed of information retrieval and reliability of scientific results verification.

No direct analogues to the study were found in accessible publications.

ArXiv preprints related to the topics of *Artificial Intelligence* (cs.AI) and *Emerging Technologies* (cs.ET) are used as examples.

MATERIALS AND METHODS

The study was based on bibliometric records exported from the arXiv database for 2021–2025 on the topics of “Artificial Intelligence” (cs.AI) and “New Technologies” (cs.ET), relevant as of December 3, 2025. The export was performed using the arXiv OAI-PMH client. Without deduplication, 126,363 records were downloaded from (cs.AI) and 4,497 from (cs.ET). No duplicates were found in the records. All records had a completed “Abstract” field.

For (cs.ET), only two records with identical English abstracts but different titles were found: ID 2404.11277; “Quantum-inspired Techniques in Tensor Networks for Industrial Contexts” and ID 2404.17645; “Técnicas Quantum-Inspired en Tensor Networks para Contextos Industriales.” The difference between them was that the first preprint was in English, while the second was in Spanish.

For (cs.AI), the share of preprints over 5 years with subsequent peer-reviewed publications is 10.9%. For (cs.ET), the share of preprints over 5 years with subsequent peer-reviewed publications is 25.6%.

Possible explanation: quite often, authors of AI algorithms and methods limit themselves to writing preprints as a detailed explanation of their work and obtaining a DOI.

The study used 46,493 multi-word terms (lowercase letters, no hyphens, lemmatized, deduplicated) found in the abstracts of cs.AI bibliometric records.

A regularly updated dictionary containing 256,000 replacement strings was used for lemmatization.

This dictionary was used for lemmatization of titles and abstracts, as well as a lexicon of multi-word terms frequently found in the abstracts of records related to the topic cs.AI, which will further be referred to as a controlled lexicon on the topic of AI.

The VOSviewer program, version 1.6.20 [9], was used to construct graphs of the co-occurrence of key terms and their clustering.

The occurrence of terms from the AI lexicon in the combined records of titles and abstracts was performed using the ugrep utility.

Note: In this work, key terms and keywords are used interchangeably.

RESULTS AND DISCUSSION

AI keyword analysis

The terms from the AI lexicon found in each of the combined title and abstract records were deduplicated, resulting in the formation of a field of key terms, which was subsequently used in a similar way to the "Index Keywords" field in Scopus bibliometric records. For brevity, this field will be referred to as "AI Keywords". Figure 1 shows the result of clustering "AI Keywords" using the VOSviewer program.

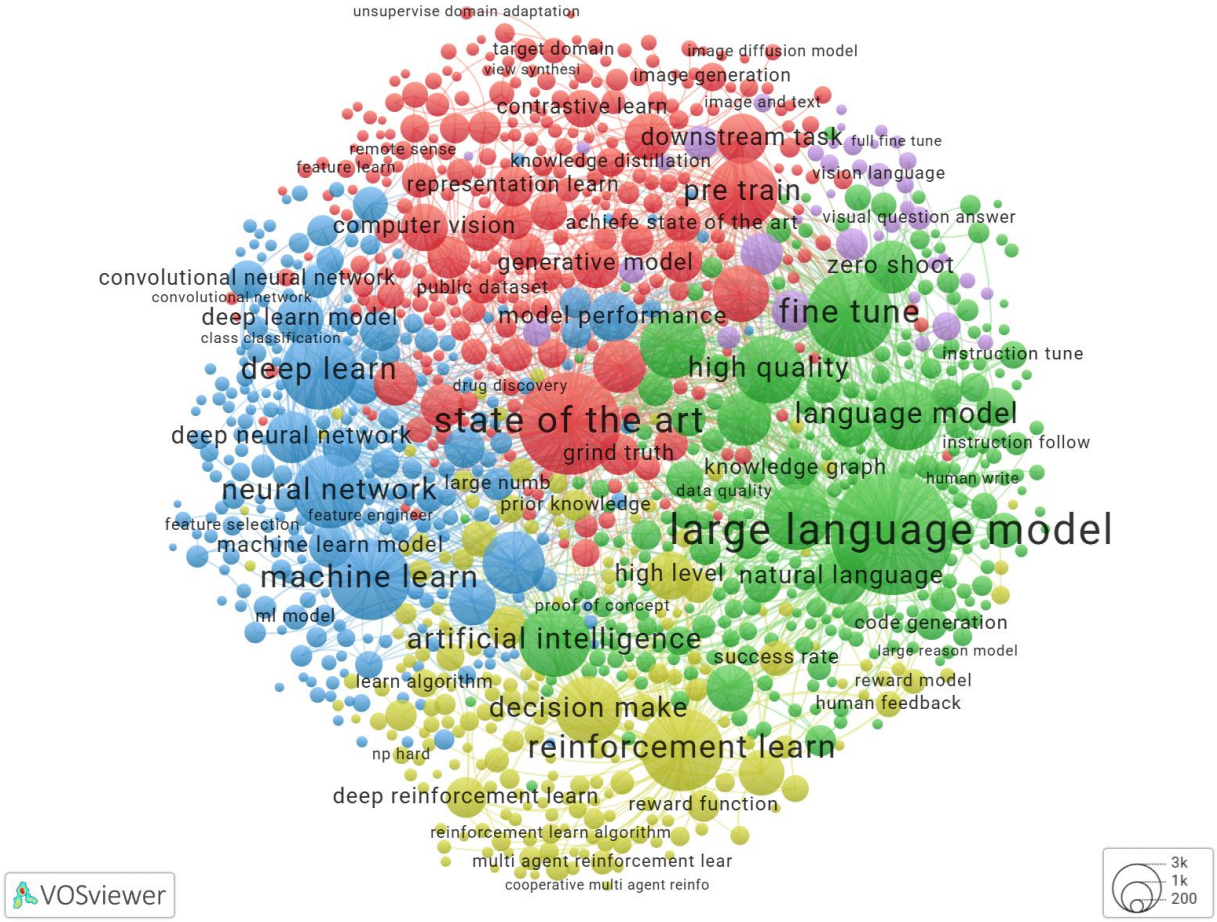

Fig. 1. Landscape of AI research based on AI Keywords co-occurrence according to ArXiv data (2021-2025)

Рис. 1. Ландшафт исследований в области ИИ на основе совместной встречаемости ключевых слов по данным ArXiv (2021-2025 гг.)

When constructing the graph, 45,263 terms were identified, of which 8,165 occurred more than 5 times. The graph was constructed based on the 1,000 terms with the highest total link strength. Five clusters were obtained. Cluster 1 – 284 terms; Cluster 2 – 254 terms; Cluster 3 – 228 terms; Cluster 4 – 195 terms; Cluster 5 – 39 terms. Thus, the four primary clusters reflect the main topics of ArXiv publications. Summary of the graph based on data from app.vosviewer.com: Items: 1000 | Links: 192,279 | Total link strength: 1,123,885 | Clusters: 5.

The most frequently occurring terms that reveal the general theme of all bibliometric records: large language model, fine tune, reinforcement learning, machine learning, deep learning, neural network, artificial intelligence, pre train, high quality, decision making. It should be noted that the figure shows lemmatized terms.

The use of a controlled lexicon on a large sample (126,363 entries) yielded a small number of clusters with clearly interpretable themes.

The terms in all tables are presented after lemmatization, in the form in which they are used in the prepared text of the title and abstracts, as well as in the lexicon used for searching.

Table 1 shows the 20 most frequently occurring terms in the first cluster (red) and their Total link strength and Occurrences.

Table 1

The 20 most frequently used terms in the first cluster, revealing the theme of state-of-the-art approaches used by AI

Таблица 1

20 наиболее часто используемых терминов в первом кластере, раскрывающем тему передовых подходов, используемых ИИ

label	**Total link strength**	**Occurrences**
state of the art	54054	12888
pre train	31378	6050
fine grain	17322	3961
base model	16432	3501
downstream task	15976	3116
generative model	11827	2764
computer vision	13433	2759
diffusion model	11545	2689
real world dataset	10028	2394
graph neural network	9009	2377
attention mechanism	10108	2242
multi modal	10010	2182
grind truth	8732	2105
supervise learn	10435	2083
representation learn	9498	2014
contrastive learn	8471	1732
achiefe state of the art	7520	1702
latent space	7325	1686
data augmentation	8323	1677
transfer learn	8000	1517

Considering the terms *pre train, fine grain, base model, downstream task, generative model,* a possible relevant topic could be formulated as the adaptation of a basic generative model to a specific applied task.

Examples of publications that reveal the formulated task include:

The authors of [10] introduced GraphPrompt as a new framework that integrates pre-training and prompting for graph neural networks, addressing the substantial supervision needed in supervised learning. It standardizes the pre-training and downstream tasks within a common template and utilizes a learnable prompt to effectively guide the downstream task in identifying relevant knowledge from the pre-trained model in a task-specific way.

Transfer learning is essential for training deep neural networks on new tasks. This method involves two stages: pseudo pre-training, which utilizes an artificially synthesized dataset from conditional source generative models, and pseudo semi-supervised learning, which employs semi-supervised learning algorithms on labeled target data and generated unlabeled pseudo samples created by combining the source classifier with generative models [11].

Table 2 shows the 20 most frequently occurring terms in the second cluster (green) and their Total link strength and Occurrences.

Table 2

Top 20 terms in the second cluster, revealing the topic of large language models

Таблица 2

20 наиболее частотных терминов во втором кластере, раскрывающих тему больших языковых моделей

label	**Total link strength**	**Occurrences**
large language model	69504	18639
fine tune	46408	9140
artificial intelligence	22312	6318
language model	24204	5889
high quality	26150	5692
open source	23114	5389
train data	24523	5294
natural language process	17340	3609
natural language	15692	3595
zero shoot	16989	3469
ai system	9136	2717
domain specific	12109	2341
knowledge graph	8852	2175
question answer	8688	1866
generative ai	5705	1735
evaluation metric	7315	1725
retrieval augment generation	6517	1510
synthetic data	7235	1465
large model	6940	1450
code generation	5493	1271

This topic can be described using the following sequence of terms: *large language model, fine-tuning, high quality, open source, domain specificity* are the main properties required for the model. To achieve this, you will need: *training data, natural language processing, synthetic data, knowledge graph, evaluation metrics*. Next are the areas of application: *AI system, questions and answers, generative AI, retrieval augment generation, code generation.*

Examples of publications that reveal the stated task include:

Fine-tuning entails further training a pretrained model, such as a Large Language Model (LLM), on a custom dataset to tailor it for specific tasks. The review [12] discusses key methodological strategies for fine-tuning LLMs, outlines general steps for the process, and presents specific use cases within medical subspecialties to demonstrate these approaches.

The advancement of Large Language Models for specific applications in materials science and engineering relies on fine-tuning strategies tailored for technical capabilities. The study [13] investigates the impacts of Continued Pretraining, Supervised Fine-Tuning, and preference-based optimization methods like Direct Preference Optimization and Odds Ratio Preference Optimization on LLM performance. Results indicate that merging multiple fine-tuned models can produce capabilities beyond those of individual models, highlighting model merging as a transformative process that fosters significant advancements through nonlinear interactions between model parameters.

Table 3 shows the 20 most frequently occurring terms in the third cluster (blue) and their Total link strength and Occurrences.

Table 3

Top 20 terms in the third cluster, revealing the topic of machine learning

Таблица 3

Топ-20 терминов в третьем кластере, раскрывающем тему машинного обучения

label	Total link strength	Occurrences
machine learn	32342	8109
deep learn	32200	7263
neural network	25772	6679
real time	20104	4786
deep neural network	13535	3276
model performance	14450	3052
deep learn model	13097	2763
black box	11796	2663
machine learn model	9644	2282
high accuracy	8932	2031
loss function	8623	1937
federate learn	7071	1893
convolutional neural network	8456	1790
image classification	7721	1494
deep learn base	6253	1401
large numb	5227	1279
anoma detection	5299	1198
model architecture	5590	1193
feature extraction	5833	1181
train dataset	5285	1085

This topic can be described using the following sequence of terms: *machine learning, deep learning, deep neural network, deep learning model, machine learning model, federated learning*, these terms reflect the subject matter itself. To implement this, a *train dataset and large numbers* are required. Implementation requires: *real time, high accuracy, loss function, and feature extraction*. The main applications are: *image classification and anomaly detection*.

If ***train data*** is used for a ***large language model***, then for ***machine learning → train dataset***.

Examples of publications that reveal the task described by the terms *machine learning, real time, deep neural network, and model performance*:

A computer vision approach for real-time object detection on low-power devices is both economically appealing and technically challenging. The paper [14] presents benchmark results on popular deep neural network models, offering insights into the trade-offs among accuracy, speed, and computational efficiency.

The paper [15] discusses benchmarks for popular deep neural network models, focusing on trade-offs among accuracy, speed, and computational efficiency crucial for real-time operation in production. It highlights the impact of inference latency, throughput, and resource utilization on user experience, service reliability, and operational costs. The paper emphasizes the necessity of real-time monitoring to address inconsistencies in input data and workload fluctuations, ensuring efficiency and integrity across various deployment environments, including cloud-hosted services, data centers, and edge devices.

Table 4 shows the 20 most frequently occurring terms in the fourth cluster (khaki) and their Total link strength and Occurrences.

Table 4

Top 20 terms in the fourth cluster, revealing the topic of reinforcement learning

Таблица 4

Топ-20 терминов в четвертом кластере, раскрывающем тему обучения с подкреплением

label	**Total link strength**	**Occurrences**
reinforcement learn	33978	8374
decision make	22983	5526
high level	12572	2882
multi agent	10521	2581
deep reinforcement learn	7986	2043
model base	8515	1940
autonomous drive	7448	1578
success rate	6232	1516
reward function	5853	1284
optimization problem	4476	1215
low level	5664	1214
prior knowledge	5333	1199
real world data	4255	987
learn algorithm	3592	969
autonomous vehicle	4078	968
multi agent system	3544	894
imitation learn	3538	858
human feedback	4353	843
reward model	3793	760
learn framework	3151	742

Deep *reinforcement learning* is a highly sought-after topic, especially in tasks such as *decision making, multi-agent systems, autonomous driving, and autonomous vehicles*. To achieve this, it is necessary to have: *real-world data, prior knowledge, imitation learning, human feedback, and a learning framework.*

Examples of publications that reveal the task described by the terms *deep reinforcement learning, decision making, and multi-agent system:*

The article [16] aims to provide an overview of current multiagent deep reinforcement learning (MDRL) literature, revisiting key components adapted from MAL and RL. It also offers guidelines for new practitioners, discusses lessons learned, recent benchmarks, and outlines research opportunities. Additionally, it critically addresses practical challenges associated with MDRL, such as implementation and computational requirements.

The article [17] discusses the significance of multiagent deep reinforcement learning (MADRL) in scenarios where multiple agents must communicate and collaborate to tackle complex tasks. It surveys various approaches to addressing challenges in MADRL, including nonstationarity, partial observability, continuous state and action spaces, multiagent training schemes, and transfer learning. The merits and drawbacks of these methods are analyzed, along with their applications.

The fifth cluster is the smallest, so we will only provide a list of the main terms describing it: *foundation model, task specific, pre-train model, vision language model, multimodal large language model, continual learning, catastrophic forgetting, vision language, visual question answering, mixture of experts, image captioning, image and text, parameter efficient, parameter efficient fine-tuning, large vision language model, low rank adaptation, vision model, vision and language, large multimodal model, multimodal LLM*. The most interesting topics seem to be *vision language model, large vision language model, image and text*.

The use of a controlled thematic lexicon in combination with extensive bibliometric records leads to the formation of a limited number of easily interpretable clusters. During text analysis, spelling errors or new terms are inevitably identified, but since the lexicon is a regular text file, it is easy to make changes to it.

ArXiv has its own categorization of research topics, which is quite compact and uses few defining terms. However, due to the large number of exported bibliometric records, it is possible to construct a network reflecting the connection between publications on the topic cs.AI and other classes. The results are shown in Figure 2.

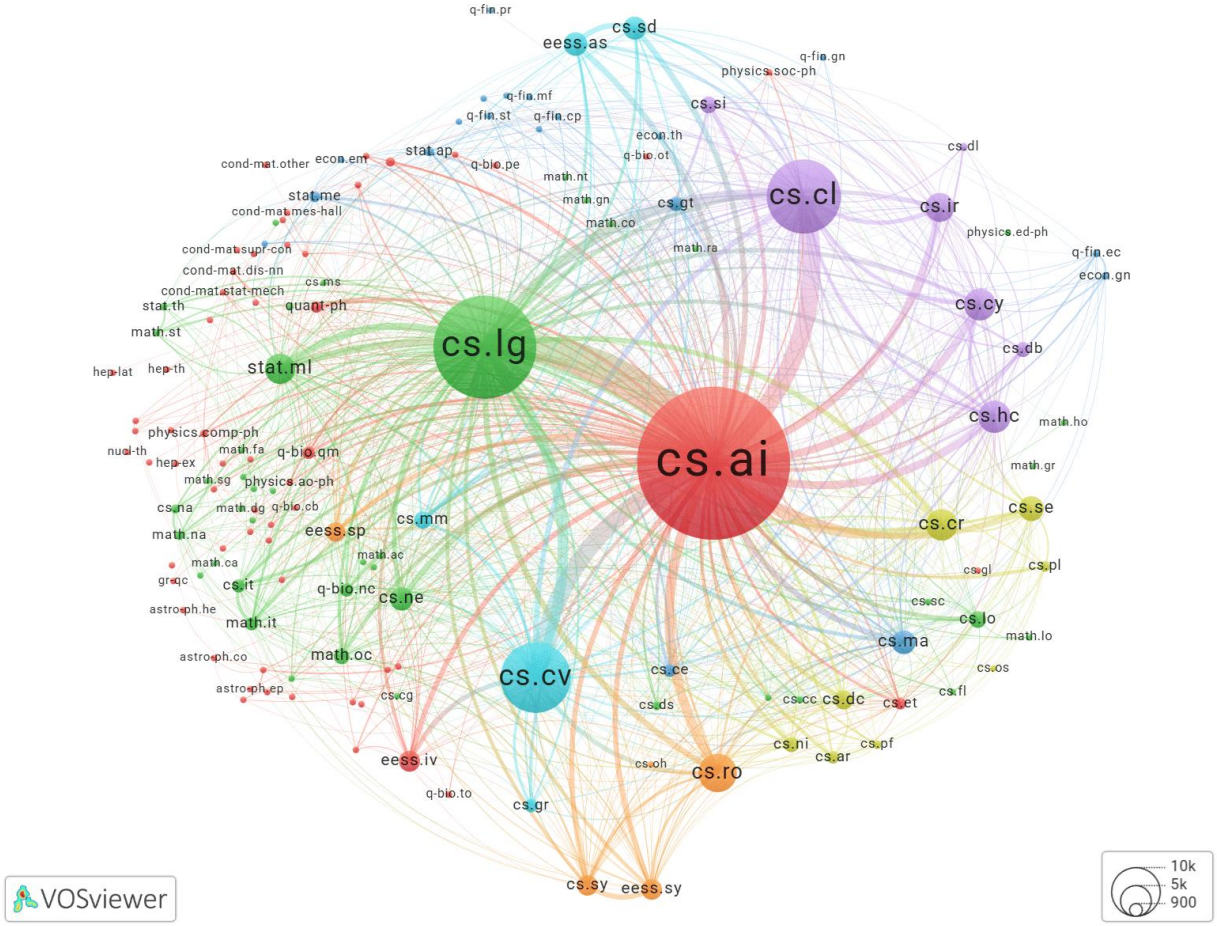

Fig. 2. Network of co-occurrence of ArXiv classes to which records on the topic cs.AI are assigned
Рис. 2. Сеть совместной встречаемости классов ArXiv, к которым отнесены записи по теме cs.AI

Here: cs.AI → Artificial Intelligence; cs.CV → Computer Vision and Pattern Recognition; cs.CL → Computation and Language; cs.RO → Robotics; cs.HC → Human-Computer Interaction; cs.IR → Information Retrieval; cs.CR → Cryptography and Security.

For more detailed information, visit: https://arxiv.org/category_taxonomy

Summary of the graph based on data from app.vosviewer.com: Items: 145 | Links: 2944 | Total link strength: 322443 | Clusters: 7

All variants of the VOSviewer layout algorithms were tested for best readability in network visualization, and the LinLog method was chosen to improve the distinguishability of node clusters and clarity between clusters.

ET keyword analysis

Terms from the AI lexicon found in each of the combined records of titles and abstracts of bibliometric records belonging to the cs.ET class and formed as “ET Keywords.” Figure 3 shows the result of clustering “ET Keywords” using the VOSviewer program.

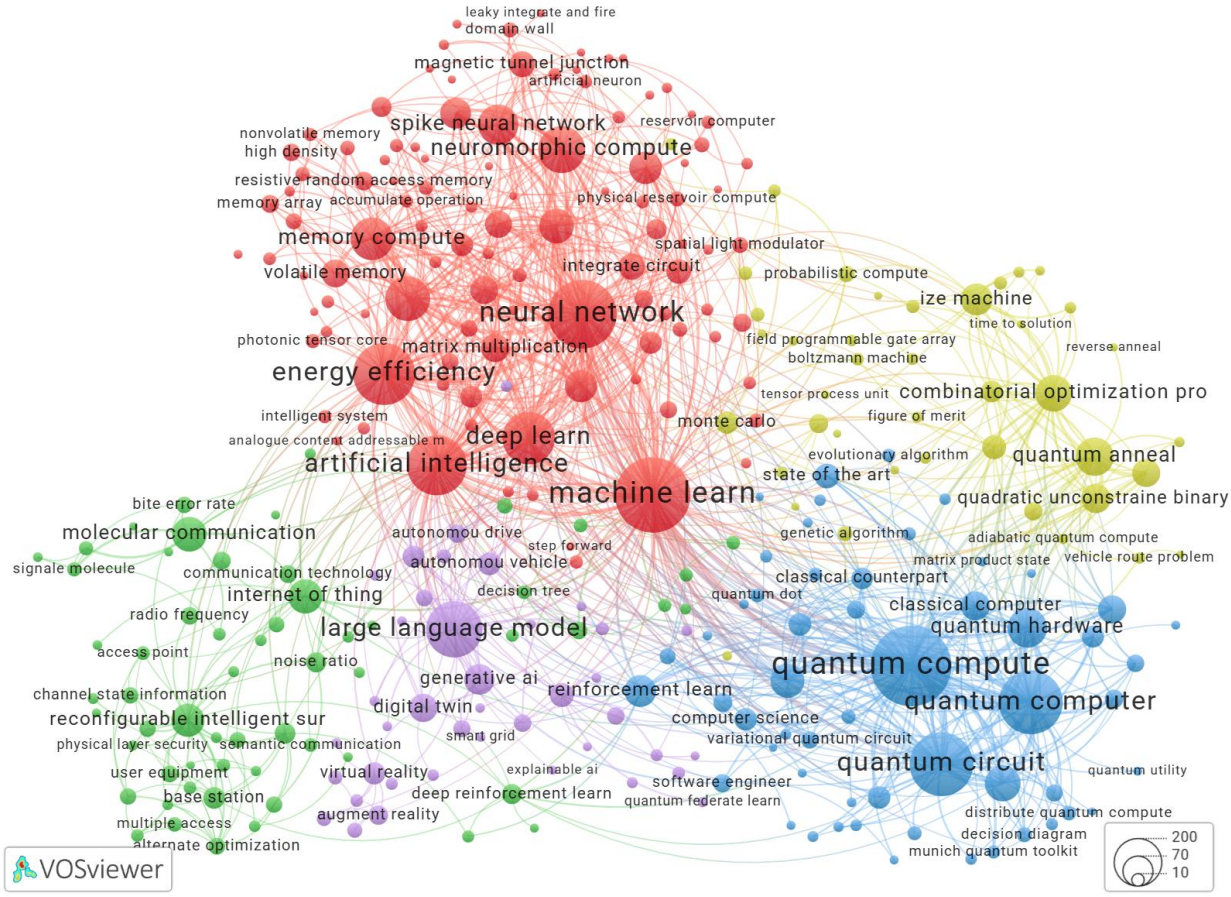

Fig. 3. Landscape of research topics in the field of Emerging Technologies based on the co-occurrence of ET keywords according to ArXiv data (2021–2025)

Рис. 3. Ландшафт исследовательских тем в области перспективных технологий на основе совместной встречаемости ключевых слов ET по данным ArXiv (2021–2025 гг.)

Summary of the graph based on data from app.vosviewer.com: Items: 300 | Links: 4293 | Total link strength: 10805 | Clusters: 5

A significant difference between Emerging Technologies and AI are clusters: #3 with terms: *quantum computing, quantum computer, quantum circuit, quantum hardware, quantum computation, quantum machine learning*; #4 with terms: *quantum annealing, combinatorial optimization problem, Ising Machine, quadratic unconstrained binary optimization, quantum annealers, combinatorial optimization, Monte Carlo*; and #1, in which the term *energy efficiency* appears among the terms *machine learning, neural network, artificial intelligence, deep learning, neuromorphic computing, and deep neural network.* The latter may indicate that the methods listed are important for *energy efficiency* and that *energy* is important for implementing these methods.

Table 5 shows the 20 most frequently occurring terms in the first cluster (red) and their Total link strength and Occurrences.

Table 5

Top 20 terms in the first cluster (red) reflecting the current tasks of Emerging Technologies

Таблица 5

20 наиболее часто встречающихся терминов в первом кластере (красный цвет), отражающих текущие задачи в области перспективных технологий

label	**Total link strength**	**Occurrences**
machine learn	1138	425
neural network	928	352
energy efficiency	720	288
artificial intelligence	583	268
deep learn	510	207
neuromorphic compute	399	165
deep neural network	366	150
memory compute	364	136
spike neural network	263	121
artificial neural network	252	90
reservoir compute	158	79
convolutional neural network	219	78
neuromorphic hardware	181	71
matrix multiplication	173	60
optical neural network	159	59
volatile memory	126	58
photonic neural network	143	55
integrate circuit	95	52
magnetic tunnel junction	128	50
activation function	144	45

Energy efficiency is a crucial term within the context of machine learning, neural networks, and artificial intelligence, highlighting the significant energy costs associated with these technologies.

This topic covers various aspects such as energy efficiency, optimizing energy networks, and data center energy use for AI, supported by relevant publications.

The research [18] examines the application of Machine Learning and AI in enhancing energy efficiency, predicting energy consumption trends, and optimizing energy systems in the USA. Utilizing datasets related to household energy usage, electric vehicle trends, and smart grid analytics, the study employs advanced techniques like deep learning, regression models, and ensemble learning to improve forecasting accuracy for better resource allocation.

The integration of Artificial Intelligence into energy systems enhances smart grid infrastructures by automating processes, improving demand forecasting, and optimizing grids. However, this combination introduces significant challenges concerning safety and security. The paper [19] analyzes the cybersecurity implications of AI in smart grids, focusing on the role of IoT systems and decentralized energy resources, while evaluating existing cybersecurity frameworks to identify vulnerabilities within an AI-enabled grid ecosystem.

Large language models (LLMs) are transforming technology and daily life, driven by extensive training on vast datasets. However, the significant electricity demand for their training and inference presents a critical challenge. The review paper [20] examines the LLM lifecycle, focusing on estimating electricity consumption and carbon emissions using statistical data.

Table 6 shows the 20 most frequently occurring terms in the second cluster (green) and their Total link strength and Occurrences.

Table 6

Top 20 terms in the second cluster (green) reflecting the current tasks of Emerging Technologies

Таблица 6

20 наиболее часто встречающихся терминов во втором кластере (зеленый цвет), отражающих текущие задачи в области перспективных технологий

label	**Total link strength**	**Occurrences**
molecular communication	96	90
internet of thing	135	87
reconfigurable intelligent surface	152	82
base station	90	35
unman aerial vehicle	77	32
deep reinforcement learn	92	30
noise ratio	76	30
communication technology	31	24
energy harvest	43	24
radio frequency	48	23
random forest	64	22
alternate optimization	56	21
user equipment	42	21
bite error rate	52	20
channel estimation	42	20
channel state information	38	20
integrate sense and communication	42	20
intelligent reflect surface	40	20
spectral efficiency	48	19
chemical reaction network	20	18

The *reconfigurable intelligent surface* ranks third in frequency, but has the highest Total link strength, as does the Internet of Things. The Internet of Things, reconfigurable intelligent surfaces, and base stations are important in communication technology, involving energy harvesting, radio frequency, and channel state information.

More details on this task can be found in [21], which offers a review and categorization of current RIS positioning research, along with insights into future directions. Reconfigurable intelligent surfaces (RIS) are proposed as a potential technology for 6G wireless communication, particularly in integration with IoT for positioning applications.

Another topic worthy of attention may be the one described by the terms: *molecular communication* and *chemical reaction network*.

Molecular communication aims to transmit information using molecules, while *chemical reaction networks* offer models and methods for this.

Table 7 shows the 20 most frequently occurring terms in the third cluster (blue) and their Total link strength and Occurrences.

Table 7

Top 20 terms in the third cluster (blue) reflecting the current tasks of Emerging Technologies

Таблица 7

20 наиболее часто встречающихся терминов в третьем кластере (синий цвет), отражающих текущие задачи в области новых технологий

label	**Total link strength**	**Occurrences**
quantum compute	1128	486
quantum computer	738	305
quantum circuit	653	303
quantum hardware	360	117
quantum computation	226	96

quantum machine learn	296	86
reinforcement learn	172	76
classical computer	213	63
quantum approximate optimization algorithm	134	59
state of the art	103	46
quantum neural network	151	39
computer science	65	38
quantum error correction	93	38
variational quantum algorithm	105	36
classical counterpart	117	35
search algorithm	52	27
quantum process unit	65	26
software engineer	51	25
parameterize quantum circuit	82	24
variational quantum eigensolv	60	23

The dominant theme of this cluster revolves around *quantum computing* and its application in *quantum reinforcement learning* (QRL). Key concepts *include quantum circuits, quantum hardware, and quantum machine learning*. QRL involves utilizing principles and algorithms from *quantum computing* to improve the efficiency of *classical reinforcement learning systems*. The distinct feature of quantum principles, like superposition, enables an reinforcement learning agent to simultaneously explore multiple states and actions, thereby accelerating the learning process.

A good instance is the publication [22], which highlights Quantum Reinforcement Learning through Variational Quantum Circuits, achieving a 90% reduction in trainable parameters and enhanced efficiency in reinforcement learning tasks, emphasizing QRL's scalability and suitability for complex scenarios.

Table 8 presents the 20 most frequently occurring terms in the fourth cluster (khaki) and their Total Link Strength and Occurrences.

Table 8

Top 20 terms in the fourth cluster (khaki) reflecting current tasks in Emerging Technologies

Таблица 8

20 наиболее часто встречающихся терминов в четвертом кластере (хаки), отражающих текущие задачи в области новых технологий

label	**Total link strength**	**Occurrences**
quantum anneal	269	110
combinatorial optimization problem	257	100
ize machine	155	76
quadratic unconstraine binary optimization	175	64
quantum annealers	149	59
combinatorial optimization	126	43
monte carlo	98	37
simulate anneal	121	35
proof of concept	56	27
travel salesman problem	60	25
probabilistic compute	51	20
genetic algorithm	44	17
differential equation	34	14
random numb generator	40	14
coherent ize machine	31	13
full connect	35	13
boltzmann machine	45	12
stochastic magnetic tunnel junction	38	12
boolean satisfiability	24	11
random numb generation	26	11

The main topic is devoted to solving complex combinatorial optimization problems using methods inspired by quantum mechanics or statistical physics. Key concepts include *Ising Machine, quantum annealers and Quadratic Unconstrained Binary Optimization* designed to efficiently solve problems in this area.

Examples of peer-reviewed publications that explore this topic include:

Quantum annealing (QA) effectively addresses combinatorial optimization problems, showcasing its practical use in contrast to classical simulated annealing methods through recent studies [23].

Quantum annealing demonstrates potential advantages over classical methods in addressing combinatorial optimization problems, highlighted by a comparative analysis of its strengths and limitations. The study [24] evaluates specific problems, such as the Traveling Salesman Problem and Quadratic Assignment Problem, while discussing performance metrics and scalability.

The fifth cluster reflects the topic of large language models, which is well represented in Table 2. To avoid overloading this article, it will not be discussed here.

Using exported bibliometric records related to the cs.ET class, a network was constructed reflecting the relationship between cs.ET and other classes. The results are shown in Figure 4.

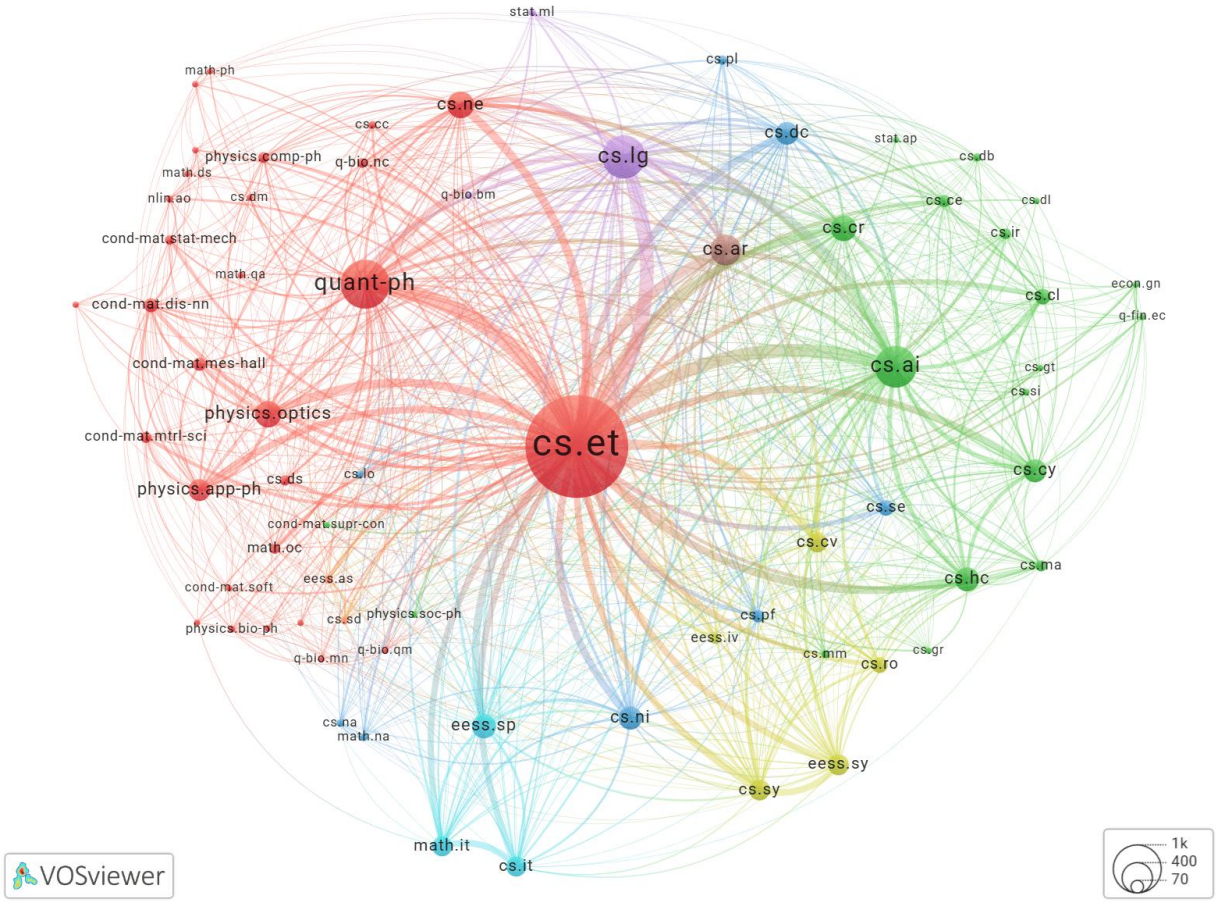

Fig. 4. Network of co-occurrence of ArXiv classes to which records on the topic cs.ET (Emerging Technologies) are assigned

Рис. 4. Сеть совместной встречаемости классов ArXiv, к которым отнесены записи по теме cs.ET (Передовые технологии)

Summary of the graph based on data from app.vosviewer.com: Items: 70 | Links: 884 | Total link strength: 13430 | Clusters: 8.

In the graph, the red cluster cs.ET, associated with the *quant-ph* class, has 1015 occurrences, while the green cluster cs.AI has 740 occurrences, indicating that cs.AI is not the most common class co-occurring with cs.ET.

Table 9 shows the main characteristics of the most common classes presented in Figure 4.

Table 9

Five most frequently occurring classes in records related to cs.ET

Таблица 9

Пять наиболее часто встречающихся классов в записях, относящихся к cs.ET

label	Total link strength	Occurrences	Avg. pub. year	
cs.et	7693	4497	2023.4505	
quant-ph	1768	1015	2022.8108	
cs.lg	2130	794	2023.8275	
cs.ai	*2102*	*740*	*2024.2689*	
cs.ar	855	406	2023.4433	

In the context of cs.ET, cs.AI topics rank only third after quant-ph and cs.LG (Machine Learning), but they appear more frequently in new publications: Avg. pub. year = 2024.2689. And Total Link Strength for cs.AI is greater than that for quant-ph.

Note: quant-ph – Quantum Physics; cs.LG – Machine Learning; cs.AR – Hardware Architecture.

The intersection of cs.ET and cs.AI topics can be effectively defined through the use of preprint metadata, which offers public and rapid access. For selecting trustworthy sources as research examples, it is advisable to concentrate on publications found in highly ranked journals.

Data obtained can identify relevant publications in peer-reviewed journals. The sample includes works related to cs.ET and cs.AI, with the ArXiv platform allowing retrieval of DOIs for publications where preprint materials are posted.

As an example, let us consider the topic of decision-making and its implementation in the fields of cs.ET and cs.AI. Analysis of exported bibliometric data shows that such work could be the article [25], in which "The authors developed Yodeai to enhance knowledge work processes, enabling information professionals to explore and synthesize unstructured data, such as product reviews and customer interview transcripts, through interactive widgets–customizable templates that convert unstructured data into structured insights".

Experts in their field are well acquainted with publications on their subject, but less familiar with research in related fields. A formalized approach to searching for relevant topics can facilitate the study of issues beyond familiar boundaries, in particular by transferring knowledge from other fields facing similar problems. However, experts rarely conduct comprehensive bibliometric analysis in related fields of knowledge.

SEVERAL ASPECTS CONCERNING FURTHER WORK WITH THE LEXICON

This section highlights the need for further research on the compilation and editing of controlled vocabularies. The VOSviewer program used in bibliometric analysis employs 1,000 terms by default to construct a network of related terms. First, terms that appear 5 or more times are selected, and from these, 1,000 terms with the highest overall connection strength are picked. Analysis of the cs.AI sample shows that of these 1,000 terms, 29 consist of 5 words, 43 consist of 4 words, 247 consist of 3 words, 659 consist of 2 words, and indicates 22 terms that should be excluded from the lexicon, such as *state of the art large language model*.

Given that there are only a few terms containing five words, it is advisable to manually save only those that are relevant to the research being conducted, such as: *pre train large language model, multi modal large language model, fine tune large language model, cooperative multi agent reinforcement learning, multi agent deep reinforcement learning*. Such a change in the lexicon is largely subjective, but subjectivity is unavoidable when analyzing text to identify relevant topics for research. The same applies to terms consisting of 4 and fewer terms, but this is a more extensive task.

Further research should analyze the sequence of searching for multi-word terms. In this work, the search for the longest substrings was implemented first, followed by shorter ones. However, a more accurate

method should take into account the “shielding” effect of words in complex terms. For example, in a situation where the phrase “supervise fine tune large language model” is present in the text, but there is no corresponding long term in the lexicon, the term “fine tune large language model” may overshadow the match for “supervise fine tune.”

Another issue not addressed by this research method is that repeated terms can be important, especially when comparing abstracts using the weighted Jaccard similarity coefficient. In abstracts, the abbreviation is often written with its full form the first time it appears, and then only as an abbreviation thereafter. However, abbreviations can have different meanings in a set of texts, for example, A2A → agent to agent; all-to-all; agriculture to agriculture; or AA → autonomous agent; authorship attribution; actuation attacks; adversarial attacks; anomaly assumption; archetypal analysis. Therefore, abbreviations must be replaced separately in each abstract.

CONCLUSIONS

The study shows that using a controlled lexicon of multi-word terms from titles and abstracts, as an analogue to Scopus Index keywords, yields easily interpretable results. This highlights 4–5 clusters of co-occurring terms that effectively describe understandable topics. This applies to both a large sample of 126,363 bibliometric records and an average sample of 4,497 records. The study results suggest focusing on terms consisting of 2-4 words to achieve optimal results.

Updating and editing a controlled lexicon is the most time-consuming process, primarily because it is difficult to formalize as it is quite subjective. This subjectivity affects the prioritization of relevant research topics. The need to precisely define key terms arises from the absence of keywords in many exported records from high-quality sources of bibliometric records, such as ArXiv, OnePetro, and Dimensions.ai.

The proposed two-stage method for identifying promising research topics, which involves extracting relevant terminology from a large sample of new open-access preprints and then conducting a targeted search for related publications in highly cited journals, has shown its effectiveness. To identify relevant peer-reviewed publications, both ArXiv metadata and artificial intelligence-based search engines such as Semantic Scholar, Elicit, or ScienceOS can be used to find publications based on the terminology found in the first stage.

Among promising research topics, it is worth highlighting one that lies at the intersection of cs.AI and cs.ET: solving complex combinatorial optimization problems using physical systems inspired by quantum mechanics or statistical physics and AI. That is, the use of a hybrid algorithm in which a quantum processor or its simulation solves combinatorially complex problems, while classical AI optimizes parameters to find the global minimum. AI acts as a “parameter optimizer.” It tunes the quantum circuit to obtain increasingly accurate results, learning at each stage.

Among the tasks that need to be continued as separate studies, we can highlight the expansion and editing of the controlled lexicon for specific research purposes. Another task is to justify the choice of a reasonable sequence for searching for multi-word terms in texts, taking into account that they may contain the same words.

Explanation: Identifying relevant topics for research does not replace the work of experts, but it reduces bias in selection and simplifies the search for resources for transferring knowledge from one field to another.

References

Список литературы

1. Yuan C. et al. The Impact of Artificial Intelligence on Economic Development: A Systematic Review: The impact of artificial intelligence on economic development // ITPHSS. 2024. V. 1, No 1. P. 130–143. DOI: 10.70693/itphss.v1i1.57

2. Erduran S., Levrini O. The impact of artificial intelligence on scientific practices: an emergent area of research for science education // International Journal of Science Education. 2024. V. 46, No 18. P. 1982–1989. DOI: 10.1080/09500693.2024.2306604

3. Levy H.V. Ethical, legal, and governance dimensions of responsible research and innovation: global perspectives and challenges in emerging technologies // Law, Ethics & Technology. 2025. DOI: 10.55092/let20250012

4. Triana I. et al. Artificial Intelligence in Innovation Research a Bibliometric Perspective // 2024 3rd International Conference on Creative Communication and Innovative Technology (ICCIT). Tangerang, Indonesia: IEEE, 2024. P. 1–6. DOI: 10.1109/ICCIT62134.2024.10701113

5. Lis A. Keywords Co-occurrence Analysis of Research on Sustainable Enterprise and Sustainable Organisation // Journal of Corporate Responsibility and Leadership. 2018. V. 5, No 2. P. 47–66. DOI: 10.12775/JCRL.2018.011

6. Koştı G., Kayadibi İ. A bibliometric analysis of artificial intelligence and machine learning applications for human resource management // Futur Bus J. 2025. V. 11, No 1. P. 179. DOI: 10.1186/s43093-025-00602-x

7. Mariani M.M. et al. Artificial intelligence in innovation research: A systematic review, conceptual framework, and future research directions // Technovation. 2023. V. 122. P. 102623. DOI: 10.1016/j.technovation.2022.102623

8. Jurowetzki R. et al. The private sector is hoarding AI researchers: what implications for science? // AI & Soc. 2025. V. 40, No 5. P. 4145–4152. DOI: 10.1007/s00146-024-02171-z

9. Van Eck N.J., Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping // Scientometrics. 2010. V. 84, No 2. P. 523–538. DOI: 10.1007/s11192-009-0146-3

10. Liu Z. et al. GraphPrompt: Unifying Pre-Training and Downstream Tasks for Graph Neural Networks // Proceedings of the ACM Web Conference 2023. Austin TX USA: ACM, 2023. P. 417–428. DOI: 10.1145/3543507.3583386

11. Yamaguchi S. et al. Transfer learning with pre-trained conditional generative models // Mach Learn. 2025. V. 114, No 4. P. 96. DOI: 10.1007/s10994-025-06748-7

12. Anisuzzaman D.M. et al. Fine-Tuning Large Language Models for Specialized Use Cases // Mayo Clinic Proceedings: Digital Health. 2025. V. 3, No 1. P. 100184. DOI: 10.1016/j.mcpdig.2024.11.005

13. Lu W., Luu R.K., Buehler M.J. Fine-tuning large language models for domain adaptation: exploration of training strategies, scaling, model merging and synergistic capabilities // npj Comput Mater. 2025. V. 11, No 1. P. 84. DOI: 10.1038/s41524-025-01564-y

14. Institute of Information Technology and Intelligent Systems, Kazan Federal Universit et al. Comparative analysis of neural network models performance on low-power devices for a real-time object detection task // Computer Optics. 2024. V. 48, No 2. P. 242–252. DOI: 10.18287/2412-6179-CO-1343

15. Real-Time Performance Monitoring for Deep Learning Models in Production // International Journal of Intelligent Systems and Applications in Engineering. 2024. DOI: 10.17762/ijisae.v12i23s.7764

16. Hernandez-Leal P., Kartal B., Taylor M.E. A survey and critique of multiagent deep reinforcement learning // Auton Agent Multi-Agent Syst. 2019. V. 33, No 6. P. 750–797. DOI: 10.1007/s10458-019-09421-1

17. Nguyen T.T., Nguyen N.D., Nahavandi S. Deep Reinforcement Learning for Multiagent Systems: A Review of Challenges, Solutions, and Applications // IEEE Trans. Cybern. 2020. V. 50, No 9. P. 3826–3839. DOI: 10.1109/TCYB.2020.2977374

18. Ibeh C.V., Adegbola A. AI and Machine Learning for Sustainable Energy: Predictive Modelling, Optimization and Socioeconomic Impact in the USA // IJASRaR. 2025. V. 2, No 1. DOI: 10.22399/ijasrar.19

19. Fathy R.A. Artificial Intelligence in the Energy Sector: Regulatory Compliance, Challenges, and Cybersecurity Implications // 2025 IEEE Conference on Power Electronics and Renewable Energy (CPERE). Aswan, Egypt: IEEE, 2025. P. 1–6. DOI: 10.1109/CPERE65146.2025.11240055

20. Ji Z., Jiang M. A systematic review of electricity demand for large language models: evaluations, challenges, and solutions // Renewable and Sustainable Energy Reviews. 2026. V. 225. P. 116159. DOI: 10.1016/j.rser.2025.116159

21. Chen R. et al. Reconfigurable Intelligent Surfaces for 6G IoT Wireless Positioning: A Contemporary Survey // IEEE Internet Things J. 2022. V. 9, No 23. P. 23570–23582. DOI: 10.1109/JIOT.2022.3203890

22. Ravish R. et al. Optimization of Reinforcement Learning Using Quantum Computation // IEEE Access. 2024. V. 12. P. 179396–179417. DOI: 10.1109/access.2024.3506656

23. Heng S. et al. How to Solve Combinatorial Optimization Problems Using Real Quantum Machines: A Recent Survey // IEEE Access. 2022. V. 10. P. 120106–120121. DOI: 10.1109/ACCESS.2022.3218908

24. Panda S., Gadam H., Upadhyay A. INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING // SSRN Journal. 2025. DOI: 10.2139/ssrn.527140310

25. Yun B. et al. Generative AI in Knowledge Work: Design Implications for Data Navigation and Decision-Making // Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems. Yokohama Japan: ACM, 2025. P. 1–19. DOI: 10.1145/3706598.3713337

Чигарев Борис Николаевич, кандидат физико-математических наук, старший научный сотрудник Аналитического центра энергетической политики и безопасности, Институт проблем нефти и газа РАН, г. Москва, Россия

Chigarev Boris Nikolaevich, Candidate of Physical and Mathematical Sciences, Senior Researcher of Analytical Center for Energy Policy and Security, Institute of Oil and Gas Problems Russian Academy of Sciences, Moscow, Russia