DOI: 10.63527/1607-8829-2025-1-16-25

M.M. Korop (https://orcid.org/0009-0000-4921-3419),
A.V. Prybyla (https://orcid.org/0000-0003-4610-2857)

Yury Fedkovych Chernivtsi National University,
2 Kotsiubynsky str., Chernivtsi, 58012, Ukraine

Corresponding author: M.M. Korop, e-mail: *koropmykola@gmail.com*

Application of LLM to Search and Systematize the Properties of Thermoelectric Materials in Scientific Literature

Thermoelectric materials find applications in a variety of fields due to their ability to directly convert heat into electricity. Selecting the optimal thermoelectric material is a challenging task, limited by empirical, time, and economic factors. Recent advances in artificial intelligence (AI), in particular large language models (LLMs), demonstrate significant potential for automatically extracting and organizing information from the scientific literature on the properties of thermoelectric materials. This review analyzes the evolution of machine learning-based methods, from early unsupervised NLP models such as Word2Vec to modern approaches using GPT models. The research results show that LLMs allow for the efficient identification of new promising thermoelectric materials, automation of experimental data collection processes, and the formation of structured databases, which significantly accelerates the search for materials with high efficiency rates. The paper also outlines directions for further research, such as extending the methods to tabular and graphical data, as well as optimizing computational resources.

Key words: thermoelectricity, materials science, machine learning, large language models, thermoelectric energy converters, computer simulation.

Introduction

Thermoelectric materials are widely used in devices for solving applied problems in various fields, namely: powering sensors, spacecraft, cooling electronics, regulating the temperature of functional elements of medical devices, heat pumps, as well as in military equipment [1-4]. To achieve optimal operating modes of such devices, it is necessary to ensure not only the maximum value of the Joffe figure of merit, but also compliance with other efficiency criteria. In particular, such a criterion is the coefficient of economic feasibility of a thermoelectric generator proposed by Anatychuk L.I. [5], which is calculated by formula 1.

Citation: M.M. Korop, A.V. Prybyla (2025). Application of LLM to Search and Systematize the Properties of Thermoelectric Materials in Scientific Literature. *Journal of Thermoelectricity*, (1), 16 – 25. https://doi.org/10.63527/1607-8829-2025-1-16-25

$$A = \frac{mN}{S_0} \tag{1}$$

where S_0 is the cost of the generator, N is the operating time, m is the value of electricity for the country of application.

Finding the most suitable material is a complex task, limited by an empirical approach, time and economic factors, which limits the pace of technology development, as well as approaches to improve the accuracy and speed of measurements of the properties of thermoelectric materials [6-9].

New approaches based on machine learning methods for analyzing and summarizing scientific data have led to a significant intensification of research into their possible application in thermoelectricity [10-11]. To begin the widespread use of such approaches in thermoelectricity, it is necessary to form a sufficient and reliable database of the properties of thermoelectric materials. Accumulating such a database using traditional experimental methods is a costly and time-consuming process.

The scientific literature contains data obtained from decades of experimental and computational research into the properties of materials, but much of this knowledge is hidden in unstructured texts. Manually collecting thermoelectric (TE) data from thousands of papers is impractical, so there is a need to use artificial intelligence (AI) and natural language processing (NLP) to automate the collection of information. In recent years, Large Language Models (LLMs) – deep neural networks trained on large text datasets – have become powerful tools for analyzing and understanding texts. Using large language models, algorithms can analyze, process, and generate text, finding the right information in unstructured data, making them an effective tool for automated processing of scientific literature and searching for relevant knowledge.

Therefore, we set the task to consider the evolution of methods for searching and collecting information on thermoelectric materials from literary sources based on artificial intelligence: from early NLP approaches to modern LLM-based systems.

The purpose of the work is to study the efficiency of using large language models (LLM) for accumulation and systematization of data on the properties of thermoelectric materials from scientific literature, as well as the formation of a list of parameters that can be obtained as a result of this process.

1. Applying unsupervised machine learning to search for material properties in the literature

Fig.1 shows the operation diagram of the transformer, on the basis of which large speech models (LLM) operate [12]. The transformer consists of two main blocks – an encoder and a decoder – each of which contains N homogeneous layers. At the input, the sequence is first converted into fixed-dimensional vectors using embedding layers, to which positional encoding is added to preserve information about the order of the elements. Each encoder layer includes a Multi-Head Self-Attention mechanism, which allows the model to consider the context of the entire input sequence in parallel in different "heads", following which the results are passed through a layered normalization layer (Add & Norm) and a position-independent two-layer feed-forward (Feed-Forward + Add & Norm).

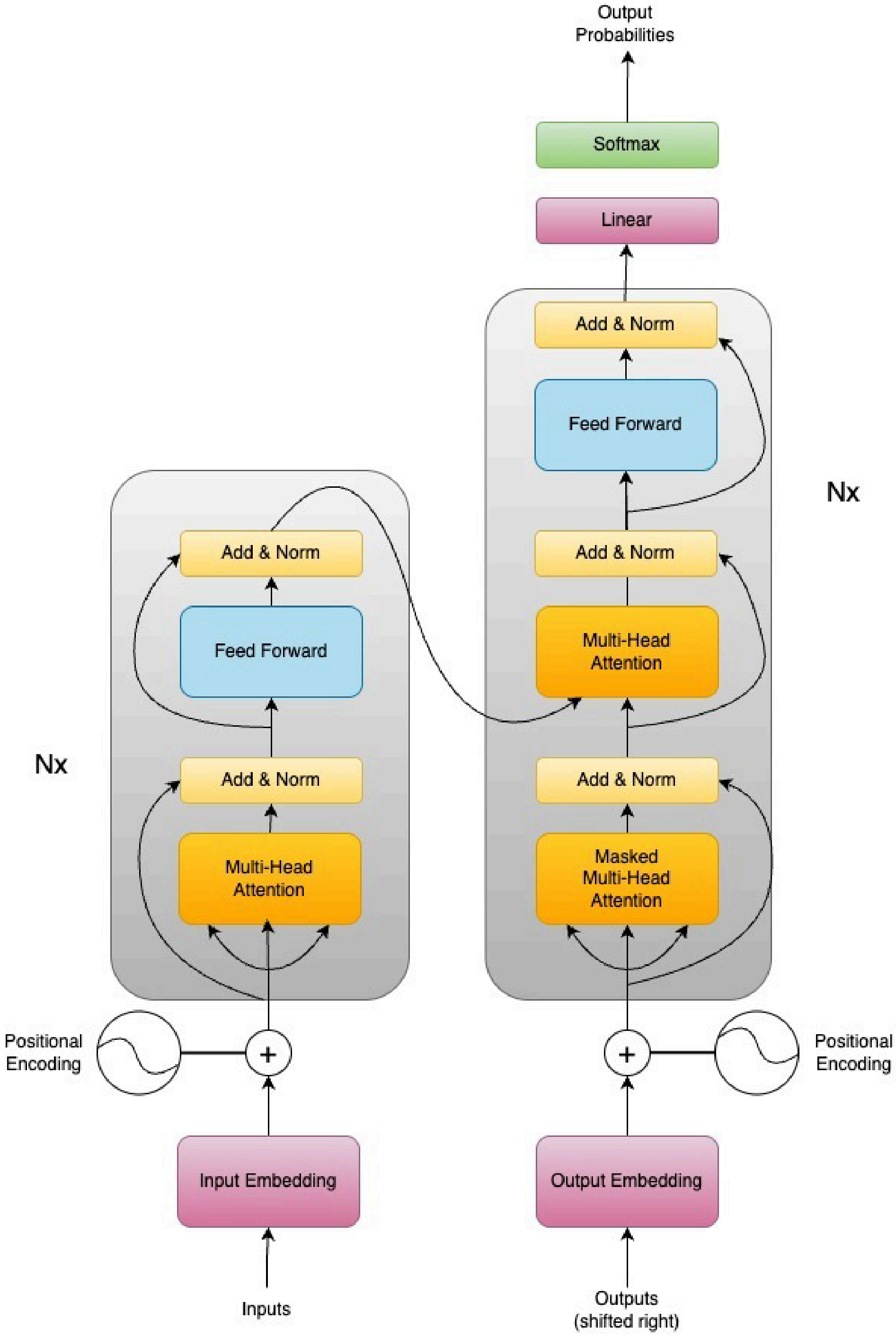

Fig. 1. Transformer operation diagram [8]

The decoder is built similarly, but contains an initial element of "masked" self-attention (Masked Multi-Head Attention), which prohibits access to "future" tokens during generation, as well as a phase of interaction with the encoder's output representations (Multi-Head Attention over encoder keys and values). After next normalization and processing through the Feed-Forward layer, the decoder output is projected through the linear layer and converted into a probability distribution over the dictionary using Softmax. This architecture provides high parallelization of learning and efficient modeling of long-term dependencies without the use of recurrent or convolution networks.

Tshitoyan V. et al. [13] were among the first to demonstrate the potential of unsupervised machine learning to uncover “hidden” knowledge about the properties of thermoelectric materials in the literature. Tshitoyan et al. trained the Word2Vec model (a two-layer neural network for word embedding) on 3.3 million abstracts from the scientific literature. For this purpose, the skip-gram algorithm of the Word2Vec model was chosen, which is capable of capturing semantic connections by predicting contextual words, which allowed obtaining 200-dimensional vector representations of each term without any manual labeling. This model was unexpectedly able to learn scientific concepts: for example, the vector for the word “iron” was closer to the word “steel” than to “organic,” reflecting fundamental chemical patterns. In the context of thermoelectric materials, the embedding of the word "thermoelectric" showed high cosine similarity with the names of certain materials (e.g., Bi_2Te_3), even though these materials were not explicitly labeled as thermoelectric in the corresponding texts. The authors interpreted such results as predictions of new potential thermoelectric materials. To confirm this idea, Tshitoyan et al. compared the materials recommended by the model with an external database containing about 48,000 compounds with calculated power factors (*ZT* factor) using density functional theory. It turned out that 7663 compounds described in the literature were not explicitly related to thermoelectric terms. When ranked by similarity to the word "thermoelectric", the top 10 candidates had high theoretical power factors (average value ~ 40.8 $\mu Wcm^{-1}K^{-2}$) – significantly higher than the known average values (~ 17.0 $\mu Wcm^{-1}K^{-2}$). This means that the model successfully identified materials with thermoelectric potential that had not been investigated. Further retrospective analysis showed that many of the materials recommended by the model were later described as thermoelectric. This work [8], published in 2019, became fundamental, proving that “hidden” knowledge from literary sources can be used to discover new materials.

While word embeddings reflect abstract relationships, other works have focused on directly extracting numerical data from texts. One of the first to create an automated database of thermoelectric materials was Sierepeklis O., Cole J. [14]. They used ChemDataExtractor 2.0, an open-source NLP tool specialized in chemical texts, adapting it for articles on thermoelectricity. They processed the texts of 60.843 articles, obtaining 22.805 data records for 10.641 unique chemical compounds, including the Seebeck coefficient, electrical and thermal conductivity, power factor, and *ZT*. The database included both experimental and theoretical results, and an accuracy of 82.25 % was achieved, making the database a valuable resource for scientists, despite some limitations (for example, the difficulty of determining the experimental or theoretical origin of the data). The work became an important example of how targeted NLP accelerates data aggregation in materials science.

By 2023, there are works by scientists describing the first more complex methods of artificial intelligence, including semi-supervised learning and the first attempts to apply large language models (LLM) in chemistry. One important study was the application of machine learning to classify new thermoelectric candidates in scientific texts. Jia X. et al. (2023) [15] presented a Positive-Unlabeled (PU) learning approach – a form of semi-supervised learning – to analyze publications to identify materials that are potentially thermoelectric. The choice of PU-learning is due to the fact that it is quite easy to collect a list of known thermoelectric compounds in the literature (the "positive" examples), while all other materials are unlabeled and not explicitly negative. The PU-classifier is able to distinguish between positive and unlabeled data, assuming that most unlabeled examples are negative, without the need for an exhaustive list of non-thermoelectric materials.

In the method proposed by Jia X. et al., articles were initially automatically marked as positive if the title mentioned the formula of a known thermoelectric material. This allowed for the creation of a training sample of validated thermoelectric articles, as opposed to a large number of unmarked publications. Then, a classifier (presumably a text model or a word frequency model) was trained to determine which unlabeled entries were indeed related to thermoelectrics. After extensive searching, the model identified 40 candidate materials that were not previously known to be of interest for applications.

To test the AI-selected candidates, the researchers performed first-principles calculations of the transport properties of each material. Strikingly, they found 20 materials (8 *p*-type and 12 *n*-type) with a theoretically predicted $ZT > 1$, the threshold for excellent thermoelectric performance. Among these candidates were entire new families of compounds, such as certain binary compounds of the AX_2 type, ternary compounds $(Cd/Zn)(GaTe_2)_2$, and quaternary chalcogenides (e.g., $CsDy_2Ag_3Te_5$), which deserved further experimental investigation. This workflow – text analysis to select candidates followed by quantum calculations – is an example of a practical application of AI: rapidly isolating promising materials from a vast chemical space.

The strong point of the PU approach model was that it used minimal expert input (only known positive examples) to efficiently exploit the literature for material discovery. However, its disadvantage was that it relied only on article titles (for labeling), which could result in important details contained in the full text being missed, and some materials could go unnoticed if their names or formulas did not clearly appear in known lists of thermoelectric materials. Nevertheless, the successful identification of high *ZT* candidates demonstrates that even partial textual input combined with semi-supervised AI can accelerate material discovery.

Significant progress has been made with the advent of GPT-3.5 (ChatGPT), a large language model with conversational capabilities and extensive knowledge. Unlike the well-defined templates of ChemDataExtractor, a large language model can, in principle, interpret sentences about a material and its properties in much the same way as a human scientist would. Early experiments looked promising: for example, there were reports that even without training on specialized materials science data, ChatGPT could identify material–property pairs, provided the queries were formulated correctly. However, challenges such as numerical accuracy, consistency, and the risk of "hallucinations" (making up facts) still needed to be overcome to trust LLM in scientific research. By the end of 2023, the foundation was laid for the dominance of information extraction techniques based on large language models, leading to rapid development of these areas in 2024.

2. Thermoelectric property collection and classification systems based on GPT models

Thway M. et al. (2024) [16] published a study in which they used the GPT-3.5 model (a 175 billion OpenAI model derived from GPT-3) to automatically extract information about the synthesis of thermoelectric materials by the solid-state method. They focused on ternary chalcogenides – compounds important for modern thermoelectrics – and sought to automatically obtain synthesis "recipes" (starting materials, annealing temperatures and durations, alloying element concentrations, etc.) from scientific papers. GPT-3.5 was chosen for its powerful natural language understanding and generation capabilities, which allowed the authors to use prompt engineering instead of manually creating information extraction rules. The authors created a reference dataset ("Gold Standard") annotated by experts and used it to evaluate the model's performance and optimize query formulation. As a result of iterative testing, an accuracy of about 73% in extracting synthesis parameters was achieved, and the model successfully obtained synthesis details from 61 out of 168 articles on chalcogenides. The study demonstrated that even without additional training, carefully formulated queries to LLM can effectively structure experimental data. The main finding was that providing multiple examples in the query significantly improved the model's accuracy. Despite certain limitations (missing hidden details or difficulties with large texts due to the GPT-3.5 context limit), this work was one of the first to use LLM in materials science experimental contexts.

Polak M.P., Morgan D. [17] presented in Nature Communications the "ChatExtract" methodology, where ChatGPT (GPT-3.5) was used to extract data on material properties with a minimum number of training examples. They perceived the model as an interlocutor: through a series of carefully thought-out queries, the model consistently found and verified the "Material – Property – Value – Unit" pattern. The innovation was a multi-stage query scheme: the model first identified candidate sentences, then confirmed the presence of the desired properties, and finally provided a structured response. The authors achieved very high accuracy: 90.8 % precision and 87.7 % recall for the elastic modulus, and similar results (~ 91 % precision, ~ 84 % recall) for the critical cooling rates of metallic glasses. This showed that LLMs can be very accurate when properly managed. A particularly important methodological aspect was to allow the model to respond "no" to avoid "hallucinations." The authors also found that retaining context in the dialogue significantly improved the accuracy of the extraction.

In contrast to previous approaches that used ready-made models, Dagdelen J. et al. [18] (2024) conducted additional training (fine-tuning) of LLM. They developed a sequence-to-sequence framework that allowed "extracting" structured data (entities and their relationships) from scientific texts, training the model on several hundred labeled examples. This allowed the model to effectively recognize compound names, synthesis steps, measured properties, and connect them. The model achieved high results (F1-score ~ 0.9), demonstrating that even a small amount of training data can significantly improve the accuracy of extraction. The disadvantage is the laboriousness of creating a training dataset and the need for specialized knowledge of machine learning. However, this approach is promising for creating highly accurate "reading machines" specializing in scientific texts.

Itani S. et al. [19] (2025) developed a database called “Large Language Model-Driven Database for Thermoelectric Materials” using the GPTArticleExtractor pipeline (now with GPT-4), which allowed to “extract” an unprecedentedly large amount of information from the literature. This project was a direct continuation of the 2022 database created using ChemDataExtractor, but with the elimination of its shortcomings thanks to the advantages of GPT-4: increased accuracy and completeness of data.

The authors systematically collected about 20.000 scientific articles (DOIs) on thermoelectric materials, mostly from Elsevier journals. Using Elsevier’s API, they retrieved the full texts of the articles in XML format and converted them into plain text for further analysis. The choice of the GPT-4 model was crucial because GPT-4 has a better understanding of complex sentences and executes instructions more consistently than previous versions.

The authors developed queries for each article, probably by analyzing individual sections (experimental results, tables, supplementary materials) to detect any mention of thermoelectric properties. In automatic mode, they created a new database covering 7123 unique thermoelectric compounds with a full set of properties. In addition to standard parameters (Seebeck coefficient, electrical conductivity, thermal conductivity, power factor, *ZT*, measurement temperature), this database also includes structural information – the type of crystal structure, lattice parameters and space group of the materials.

The inclusion of structural descriptors is an important step forward, as it allows us to investigate the dependence of properties on the structure of the material (for example, the relationship of the type of crystal lattice with high *ZT*). GPT-4, with its extensive knowledge and advanced analytical capabilities, was able to correctly identify and summarize such information that simple parsers often cannot recognize if the data format is non-standard.

The resulting dataset, with thousands of compounds, each with numerous annotated properties, is one of the most comprehensive resources in the field of thermoelectrics to date. The authors emphasize that the use of LLM allowed overcoming many limitations of manual or semi-automatic data collection. For example, the model could adapt to different forms of data representation (different units, word order, wording), which was a problem for manual approaches.

After the automatic data extraction, the authors implemented scripts to standardize units of measurement. In addition, automated cross-comparison of the “extracted” information with the source text was performed, which increased the reliability of the database. Although exact accuracy figures have not yet been published, high accuracy is expected on the basis of previous experience.

An important improvement was the structuring: each database entry has a clear labeling of the context (composition, structure, properties, temperature, etc.). Thanks to this, the database has become especially useful for training machine models and analyzing relationships in materials science. For example, it is now possible to quickly find all materials with a diamond-like cubic structure and a Seebeck coefficient > 200 μV/K at 300 K, which previously required significant literature search efforts. The authors note that challenges remain for fully automated data collection, including distinguishing data from multiple materials within a single text and determining whether data is experimentally or theoretically calculated. These issues are the subject of further research and improvement.

Conclusions

1. There are solutions for generating a sufficient database of thermoelectric material properties from the scientific literature based on a combination of automatic tools with AI and verification using more specialized tools or humans.
2. Potential areas for further improvement of the technology: processing tabular and graphical elements of scientific articles, working with foreign-language data, since the vast majority of LLM models were trained on English-language literature, further optimization of search methods to reduce the required computing power.
3. It is promising to develop an AI agent capable of continuously analyzing new articles in thermoelectric materials science, which would allow scientists to gain instant access to a relevant source of information.

Authors' information

M.M. Korop – Postgraduate.
A.V. Prybyla – Cand.Sc. (Phys.-Math.).

References

1. Anatychuk L.I., Prybyla A.V. (2017). Limiting possibilities of thermoelectric liquid-liquid heat pumps. *J. Thermoelectricity, 4*, 51 – 55.
2. Rifert V., Anatychuk L., Barabash P., Solomakha A., Usenko V., Prybyla A., Sereda V. (2019). Comparative analysis of thermal distillation methods with heat pumps for long space flights. *J.Thermoelectricity*, 4, 5 – 17. Retrieved from http://jte.ite.cv.ua/index.php/jt/article/view/70
3. Anatychuk L., Lysko V., Prybyla A. (2022). Rational areas of using thermoelectric heat recuperators. *J. Thermoelectricity*, 3-4, 43 – 67. https://doi.org/10.63527/1607-8829-2022-3-4-43-67
4. Anatychuk L., Prybyla A., Korop M., Kiziuk Y., & Konstantynovych I. (2024). Thermoelectric power sources using low-grade heat: Part 1. *J. Thermoelectricity*, 1-2, 90 – 96. https://doi.org/10.63527/1607-8829-2024-1-2-90-96
5. Anatychuk L. (2020). Efficiency criterion of thermoelectric energy converters using waste heat. J.Thermoelectricity, 4, 58 – 63. Retrieved from http://jte.ite.cv.ua/index.php/jt/article/view/47
6. Anatychuk L.I., Lysko V.V., Havryliuk M.V. (2018). Ways for quality improvement in the measurement of thermoelectric material properties by the absolute method. *J.Thermoelectricity*, 2, 90 – 100.
7. Anatychuk L.I., Lysko V.V., Havryliuk M.V., Tiumentsev V.A. (2018). Automation and computerization of measurements of thermoelectric parameters of materials.

J. Thermoelectricity, 3, 80 – 88.

8. Anatychuk L.I., Lysko V.V. (2012). Investigation of the effect of radiation on the precision of thermal conductivity measurement by the absolute method. *J. Thermoelectricity*, 1, 65 – 73.
9. Anatychuk L.I., Lysko V.V. Modified Harman's method. (2012) AIP Conference Proceedings, 1449, 373 – 376. DOI: 10.1063/1.4731574.
10. Korop M.M. (2023). Machine learning in thermoelectric materials science. In: *J. Thermoelectricity*, 1, 44 – 54. Institute of Thermoelectricity. https://doi.org/10.63527/1607-8829-2023-1-44-54
11. Anatychuk L.I., Korop M.M. (2023). Application of machine learning to predict the properties of Bi_2Te_3-based thermoelectric materials. In: *J. Thermoelectricity*, 2, 59 – 71. Institute of Thermoelectricity. https://doi.org/10.63527/1607-8829-2023-2-59-71
12. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin, I. (2017). Attention Is All You Need (Version 7). arXiv. https://doi.org/10.48550/ARXIV.1706.03762
13. Tshitoyan V., Dagdelen J., Weston L., Dunn A., Rong Z., Kononova O., Persson K.A., Ceder G., & Jain A. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. In *Nature*, 571(7763), 95 – 98. Springer Science and Business Media LLC. https://doi.org/10.1038/s41586-019-1335-8
14. Sierepeklis O., Cole J.M. (2022). A thermoelectric materials database auto-generated from the scientific literature using ChemDataExtractor. In *Scientific Data* (Vol. 9, Issue 1). Springer Science and Business Media LLC. https://doi.org/10.1038/s41597-022-01752-1
15. Jia X., Yao H., Yang Z., Shi J., Yu J., Shi R., Zhang H., Cao F., Lin X., Mao J., Wang C., Zhang Q., & Liu X. (2023). Advancing thermoelectric materials discovery through semi-supervised learning and high-throughput calculations. In *Applied Physics Letters*, 23, 20. AIP Publishing. https://doi.org/10.1063/5.0175233
16. Thway M., Low A.K.Y., Khetan S., Dai H., Recatala-Gomez J., Chen A.P., Hippalgaonkar K. (2024). Harnessing GPT-3.5 for text parsing in solid-state synthesis – case study of ternary chalcogenides. In *Digital Discovery*. 3(2), 328 – 336). Royal Society of Chemistry (RSC). https://doi.org/10.1039/d3dd00202k
17. Polak M.P., Morgan D. (2024). Extracting accurate materials data from research papers with conversational language models and prompt engineering. In *Nature Communications*. 15(1), Springer Science and Business Media LLC. https://doi.org/10.1038/s41467-024-45914-8
18. Dagdelen J., Dunn A., Lee S., Walker N., Rosen A.S., Ceder G., Persson K.A., Jain A. (2024). Structured information extraction from scientific text with large language models. In *Nature Communications,* 15(1). Springer Science and Business Media LLC. https://doi.org/10.1038/s41467-024-45563-x
19. Itani S., Zhang Y., & Zang J. (2025). Large Language Model-Driven Database for Thermoelectric Materials (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2501.00564

Submitted: 27.01.2025

Короп М.М. (https://orcid.org/0009-0000-4921-3419),
Прибила А.В. (https://orcid.org/0000-0003-4610-2857)

Чернівецький національний університет імені Юрія Федьковича,
вул. Коцюбинського 2, Чернівці, 58012, Україна

Застосування LLM для пошуку та систематизації властивостей термоелектричних матеріалів із наукової літератури

Термоелектричні матеріали знаходять застосування у різноманітних сферах завдяки можливості прямого перетворення тепла в електроенергію. Вибір оптимального термоелектричного матеріалу є складним завданням, яке обмежується емпіричними, часовими та економічними факторами. Останні досягнення в галузі штучного інтелекту (ШІ), зокрема великі мовні моделі (LLMs), демонструють значний потенціал для автоматичного збору та систематизації інформації з наукової літератури про властивості термоелектричних матеріалів. Цей огляд аналізує еволюцію методів на основі машинного навчання, від ранніх некерованих NLP-моделей, таких як Word2Vec, до сучасних підходів з використанням GPT-моделей. Результати досліджень показують, що LLM дозволяють ефективно ідентифікувати нові перспективні термоелектричні матеріали, автоматизувати процеси збирання експериментальних даних і формувати структуровані бази, що значно прискорює пошук матеріалів з високими показниками ефективності. У роботі окреслені також напрямки для подальших досліджень, такі як розширення методів на табличні та графічні дані, багатомовні моделі, а також оптимізація обчислювальних ресурсів.

Ключові слова: термоелектрика, матеріалознавство, машинне навчання, великі мовні моделі, термоелектричні перетворювачі енергії, комп'ютерне моделювання.

Надійшла до редакції 27.01.2025