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

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

Автор-кореспондент: Короп М.М., e-mail: *koropmykola@gmail.com*

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.

Вступ

Термоелектричні матеріали мають широке застосування у пристроях для вирішення прикладних задач у різних сферах, а саме: живлення сенсорів, космічних апаратів, охолодження електроніки, регулювання температури функціональних елементів медичних пристроїв, теплових насосів, а також у військовій техніці [1-4]. Для досягнення оптимальних режимів роботи таких пристроїв, необхідно забезпечити не

Цитування: Короп М.М., Прибила А.В. (2025). Застосування LLM для пошуку та систематизації властивостей термоелектричних матеріалів із наукової літератури. *Термоелектрика*, (1), 16 – 25. https://doi.org/10.63527/1607-8829-2025-1-16-25

тільки максимальне значення критерію добротності Йоффе, а й відповідність іншим критеріям ефективності. Зокрема, таким критерієм є коефіцієнт економічної доцільності термоелектричного генератора запропонований Анатичуком Л.І. [5], який розраховується за формулою 1.

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

де S_0 – вартість генератора, N – час роботи, m – значення електроенергії для країни застосування.

Знаходження найбільш відповідного матеріалу є складною задачею, яка лімітується емпіричним підходом, часовими та економічними факторами, що обмежує темпи розвитку технології, а також підходами для покращення точності та швидкості вимірювань властивостей термоелектричних матеріалів [6-9].

Нові підходи на базі машинних методів аналізу та узагальнення наукових даних зумовили значну інтенсифікацію досліджень їх можливого застосування в термоелектриці [10-11]. Для початку широкого використання таких підходів у термоелектриці, потрібно сформувати достатню та надійну базу властивостей термоелектричних матеріалів. Накопичення такої бази традиційними експериментальними методами є дороговартісним і трудомістким процесом.

Наукова література містить дані, отримані внаслідок десятиліть експериментальних та обчислювальних досліджень властивостей матеріалів, проте значна частина цих знань є прихованою у неструктурованих текстах. Збирання термоелектричних (ТЕ) даних вручну із тисяч статей є непрактичним, тому з’явилася необхідність у використанні штучного інтелекту (ШІ) та обробки природної мови (NLP – Natural Language Processing) для автоматизації збору інформації. Останніми роками великі мовні моделі (LLMs – Large Language Models) – глибокі нейронні мережі, навчені на великих текстових наборах даних – стали потужними інструментами для аналізу та розуміння текстів. Використовуючи великі мовні моделі, алгоритми можуть аналізувати, обробляти та генерувати текст, знаходячи потрібну інформацію у неструктурованих даних, що робить їх ефективним інструментом для автоматизованої обробки наукової літератури та пошуку релевантних знань.

Тому, нами було поставлено завдання розглянути еволюцію методів пошуку та збору інформації про термоелектричні матеріали з літературних джерел на основі штучного інтелекту: від ранніх підходів NLP до сучасних систем на базі LLM.

Метою роботи є вивчення ефективності застосування великих мовних моделей (LLM) для накопичення та систематизації даних про властивості термоелектричних матеріалів із наукової літератури, а також формування переліку параметрів, які можуть бути отримані в результаті цього процесу.

1. Застосування некерованого машинного навчання для пошуку властивостей матеріалів у літературі

На рисунку 1 представлена схема роботи трансформера, на основі якого функціонують великі мовні моделі (LLM) [12]. Трансформер складається з двох основних

блоків – енкодера та декодера – кожен із яких містить N однорідних шарів. На вході послідовність спочатку перетворюється на вектори фіксованої розмірності за допомогою шарів вбудовування (Embedding), до яких додається позиційне кодування (Positional Encoding) для збереження інформації про порядок елементів. Кожен шар енкодера охоплює механізм багатоголової самоуваги (Multi-Head Self-Attention), який дозволяє моделі враховувати контекст усієї вхідної послідовності паралельно в різних «головах», після чого результати проходять через шар шарового нормування (Add & Norm) і позиційно-незалежний двошаровий прямий зв'язок (Feed-Forward + Add & Norm).

Декодер побудований аналогічно, але містить початковий елемент «маскованої» самоуваги (Masked Multi-Head Attention), який забороняє доступ до «майбутніх» токенів під час генерації, а також фазу взаємодії з вихідними представленнями енкодера (Multi-Head Attention над ключами та значеннями енкодера). Після чергового нормування та обробки через Feed-Forward шар вихід декодера проектується через лінійний шар та перетворюється в розподіл ймовірностей по словнику за допомогою Softmax. Така архітектура забезпечує високу паралелізацію навчання та ефективне моделювання довготривалих залежностей без використання рекурентних чи згорткових мереж.

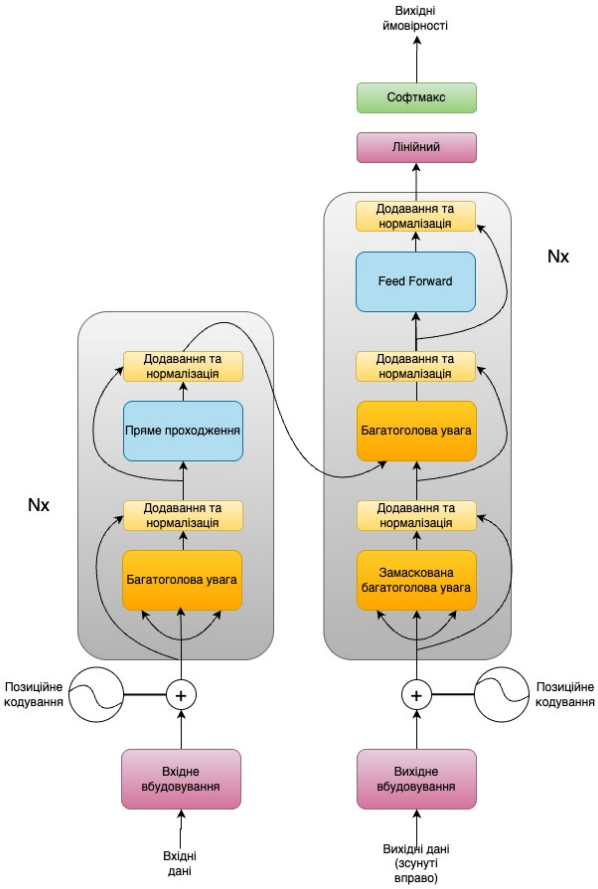

Рис. 1. Схема роботи трансформера [8]

Tshitoyan V. та ін. [13] одним із перших вдалось продемонструвати потенціал некерованого машинного навчання для виявлення «прихованих» у літературі знань про властивості термелектричних матеріалів. Tshitoyan та ін. натренували модель Word2Vec (двошарову нейронну мережу для вкладення слів) на базі 3.3 мільйонів анотацій із наукової літератури. Для цього, було обрано алгоритм skip-gram моделі Word2Vec, який здатен вловлювати семантичні зв'язки, передбачаючи контекстні слова, що дозволило отримати 200-вимірні векторні представлення кожного терміна без будь-якої ручної мітки. Ця модель несподівано змогла засвоїти наукові концепції: наприклад, вектор для слова «залізо» був ближчим до слова «сталь», ніж до «органічний», що відображає фундаментальні хімічні закономірності. У контексті термоелектричних матеріалів, вкладення слова «термоелектричний», показало високу косинусну схожість з назвами певних матеріалів (наприклад, Bi_2Te_3), навіть, якщо ці матеріали явно не позначалися, як

термоелектричні у відповідних текстах. Автори інтерпретували такі результати як прогнози нових потенційних термоелектричних матеріалів. Щоб підтвердити цю ідею, Tshitoyan та ін. порівняли матеріали, рекомендовані моделлю, із зовнішньою базою даних, що містить близько 48 000 сполук з обчисленими коефіцієнтами потужності (коефіцієнт *ZT*), за допомогою теорії функціоналу густини. Виявилось, що 7663 сполук, описаних в літературі, не були явно пов'язані з термоелектричними термінами. При ранжуванні за схожістю до слова «термоелектричний», топ-10 кандидатів мали високі теоретичні коефіцієнти потужності (середнє значення ~ 40.8 $\mu W\,см^{-1}\,K^{-2}$) – це суттєво вище за відомі середні значення (~ 17.0 $\mu W\,см^{-1}\,K^{-2}$). Це означає, що модель успішно ідентифікувала матеріали з термоелектричним потенціалом, який не було досліджено. Подальший ретроспективний аналіз показав, що багато матеріалів, рекомендованих моделлю, пізніше були описані як термоелектричні. Ця робота [8], опублікована у 2019 році, стала фундаментальною, довівши, що «приховані» знання з літературних джерел можна використовувати для відкриття нових матеріалів.

Тоді, як вкладення слів відображають абстрактні зв’язки, інші роботи були спрямовані на пряме вилучення чисельних даних із текстів. Одними з перших, хто створив автоматизовану базу даних термоелектричних матеріалів, були Sierepeklis O., Cole J. [14]. Вони використали ChemDataExtractor 2.0 – інструмент NLP з відкритим кодом, спеціалізований на хімічних текстах, адаптувавши його для статей з термоелектрики. Вони обробили тексти 60843 статей, отримавши 22805 записів даних для 10641 унікальних хімічних сполук, включаючи коефіцієнт Зеєбека, електричну та теплову провідність, коефіцієнт потужності та *ZT*. База включала в себе як експериментальні, так і теоретичні результати, вдалося досягнути точності у 82.25 %, що зробило цю базу цінним ресурсом для науковців, незважаючи на окремі обмеження (наприклад, складність визначення експериментального чи теоретичного походження даних). Робота стала важливим прикладом того, як цільовий NLP прискорює агрегування даних в матеріалознавстві.

До 2023 року існують роботи науковців, в яких описані перші більш складні методи штучного інтелекту, включно з напівкерованим навчанням та першими спробами застосування великих мовних моделей (LLM) у хімічному напрямку. Одним із важливих досліджень, було застосування машинного навчання для класифікації нових термоелектричних кандидатів у наукових текстах. Jia X. та ін. (2023) [15] Jia X. та ін. представили підхід позитивно-немаркованого (PU, Positive-Unlabeled) навчання – форму напівкерованого навчання – для аналізу публікацій з метою виявлення матеріалів, що потенційно є термоелектричними. Вибір PU-навчання обумовлений тим, що в літературі досить легко зібрати список відомих термоелектричних сполук («позитивні» приклади), тоді як усі інші матеріали є немаркованими, а не явно негативними. PU-класифікатор здатний розрізняти позитивні та немарковані дані, припускаючи, що більшість немаркованих прикладів є негативними, без необхідності мати вичерпний список не термоелектричних матеріалів.

У методі, запропонованому Jia X. та ін., спочатку статті автоматично маркувалися як позитивні, якщо у заголовку згадувалася формула відомого термоелектричного матеріалу. Це дозволило створити навчальну вибірку з підтверджених термоелектричних статей, на противагу великій кількості немаркованих публікацій. Потім, було натреновано класифікатор (імовірно, текстова модель або модель на основі частоти слів), який визначав, які немарковані записи дійсно пов'язані з термоелектриками. За результатами широкого пошуку, модель виявила 40 матеріалів-кандидатів, які раніше не були відомі як матеріали, цікаві для застосування.

Для перевірки відібраних за допомогою ШІ кандидатів, науковці провели розрахунки транспортних властивостей кожного матеріалу за допомогою методів перших принципів. Вражаюче, що серед них було виявлено 20 матеріалів (8 p-типу та 12 n-типу) з теоретично передбаченим $ZT > 1$, що є пороговим значенням для відмінних термоелектричних характеристик. Серед цих кандидатів були цілі нові родини сполук, такі як певні бінарні сполуки типу AX_2, тернарні сполуки $(Cd/Zn)(GaTe_2)_2$ та четвертинні халькогеніди (наприклад, $CsDy_2Ag_3Te_5$), які заслуговували подальших експериментальних досліджень. Цей робочий процес – текстовий аналіз для вибору кандидатів з наступними квантовими розрахунками – є прикладом практичного застосування ШІ: швидке виокремлення перспективних матеріалів із величезного хімічного простору.

Сильна сторона моделі PU-підходу полягала в тому, що вона використовувала мінімальний внесок експертів (лише відомі позитивні приклади), щоб ефективно використовувати літературу для відкриття матеріалів. Однак, недоліком було те, що вона спиралася лише на заголовки статей (для маркування), через що важливі деталі, що містяться у повному тексті, могло бути пропущено, а деякі матеріали могли залишитися непоміченими, якщо їх назви чи формули явно не фігурували у відомих списках термоелектричних матеріалів. Тим не менше, успішність у виявленні кандидатів з високим ZT свідчить про те, що навіть часткові текстові свідчення в поєднанні з напівкерованим ШІ можуть прискорити відкриття матеріалів.

Значний прогрес був забезпечений появою GPT-3.5 (ChatGPT) – великої мовної моделі з розмовними можливостями та широкими знаннями. На відміну від чітко визначених шаблонів ChemDataExtractor, велика мовна модель, в принципі, може інтерпретувати речення про матеріал і його властивості приблизно так, як це зробив би науковець-людина. Ранні експерименти виглядали перспективними: наприклад, з'явилися повідомлення про те, що навіть без навчання на спеціалізованих матеріалознавчих даних, ChatGPT міг ідентифікувати пари матеріал–властивість за умови правильного формулювання запитів. Однак такі виклики, як чисельна точність, послідовність та ризик «галюцинацій» (вигадування фактів) все ще потрібно було подолати, щоб довіряти LLM у наукових дослідженнях. До кінця 2023 року було закладено основу для домінування методик вилучення інформації на базі великих мовних моделей, що призвело до швидкого розвитку цих напрямків у 2024 році.

2. Системи збору та класифікації термоелектричних властивостей на основі моделей GPT

Thway M. та ін. (2024) [16] опублікували дослідження, у якому використали модель GPT-3.5 (175-мільярдна модель OpenAI, похідна від GPT-3) для автоматичного вилучення інформації про синтез термоелектричних матеріалів твердофазним методом. Вони зосередилися на тернарних халькогенідах – сполуках, важливих для сучасних термоелектриків – і прагнули автоматично отримувати «рецепти» синтезу (вихідні матеріали, температури і тривалість відпалу, концентрації легуючих елементів тощо) з наукових статей. GPT-3.5 було обрано за його потужні можливості розуміння і генерації природної мови, що дозволило авторам використовувати інженерію запитів (prompt engineering) замість ручного створення правил вилучення інформації. Автори створили еталонний набір даних («Gold Standard»), анотований експертами, і використовували його для оцінки ефективності моделі та оптимізації формулювання запитів. У результаті ітеративного тестування вдалося досягти точності близько 73 % у вилученні параметрів синтезу, а модель успішно отримала деталі синтезу із 61 статті зі 168 про халькогеніди. Дослідження продемонструвало, що навіть без додаткового навчання, ретельно сформульовані запити до LLM можуть ефективно структурувати експериментальні дані. Основним відкриттям стало те, що надання кількох прикладів у запиті значно підвищує точність моделі. Незважаючи на певні обмеження (пропуск прихованих деталей або труднощі з великими текстами через ліміт контексту GPT-3.5), ця робота стала однією з перших, що використала LLM у матеріалознавчих експериментальних контекстах.

Polak M.P., Morgan D. [17] представили в Nature Communications методику «ChatExtract», де ChatGPT (GPT-3.5) застосовувався для вилучення даних про властивості матеріалів із мінімальною кількістю навчальних прикладів. Вони сприймали модель як співрозмовника: через серію детально продуманих запитів модель послідовно знаходила та перевіряла шаблон «Матеріал–Властивість – Значення – Одиниця вимірювання». Інновацією стала багатоетапна схема запитів: модель спочатку ідентифікувала речення-кандидати, потім підтверджувала наявність шуканих властивостей і, нарешті, видавала структуровану відповідь. Автори досягли дуже високої точності: 90.8 % precision та 87.7 % recall для модуля пружності, та аналогічних результатів (~ 91 % precision, ~ 84 % recall) для критичних швидкостей охолодження металевих стекол. Це довело, що LLM можуть бути дуже точними при правильному керуванні запитами. Особливо важливим методологічним аспектом було дозволити моделі відповідати «ні», щоб уникнути «галюцинацій». Автори також виявили, що утримання контексту в діалозі суттєво підвищувало точність вилучення.

На відміну від попередніх підходів, що використовували готові моделі, Dagdelen J. та ін. [18] (2024) проводили додаткове навчання (fine-tuning) LLM. Вони розробили фреймворк sequence-to-sequence, який дозволяв «витягувати» структуровані дані (сутності та їх зв'язки) з наукових текстів, натренувавши модель на кількох сотнях розмічених прикладів. Це дало змогу моделі ефективно розпізнавати назви сполук, етапи з синтезу, виміряні властивості, та зв'язувати їх. Модель досягла високих результатів (F1-

бал ~0,9), демонструючи, що навіть невеликий обсяг навчальних даних може значно покращити точність вилучення. Недоліком є трудомісткість створення навчального набору даних та необхідність спеціалізованих знань з машинного навчання. Однак, такий підхід є перспективним для створення високоточних «читальних машин», що спеціалізуються на наукових текстах.

Itani S. та ін. [19] (2025) розробили базу під назвою «Large Language Model-Driven Database for Thermoelectric Materials», використовуючи конвеєр GPTArticleExtractor (тепер із GPT-4), що дозволило «витягнути» безпрецедентно великий обсяг інформації з літератури. Цей проєкт став прямим продовженням бази даних 2022 року, створеної за допомогою ChemDataExtractor, але з усуненням її недоліків завдяки перевагам GPT-4: підвищеній точності та повноті даних.

Автори систематично зібрали близько 20 000 наукових статей (DOI) про термоелектричні матеріали, переважно з журналів Elsevier. Використовуючи API Elsevier, вони отримали повні тексти статей у форматі XML і конвертували їх у простий текст для подальшого аналізу. Вибір моделі GPT-4 був вирішальним, оскільки GPT-4 володіє кращим розумінням складних речень і більш послідовно виконує інструкції порівняно з попередніми версіями.

Автори розробили запити для кожної статті, ймовірно, аналізуючи окремі секції (експериментальні результати, таблиці, додаткові матеріали), щоб виявити будь-які згадки термоелектричних властивостей. В автоматичному режимі вони створили нову базу, що охоплює 7123 унікальних термоелектричних сполук із повним набором властивостей. Окрім стандартних параметрів (коефіцієнт Зеєбека, електропровідність, теплопровідність, фактор потужності, *ZT*, температура вимірювання), ця база включає також структурну інформацію – тип кристалічної структури, параметри ґратки та просторову групу матеріалів.

Включення структурних дескрипторів є важливим кроком уперед, адже це дозволяє досліджувати залежності властивостей від структури матеріалу (наприклад, зв'язок типу кристалічної ґратки з високим *ZT*). GPT-4 з його широкими знаннями та розвинутими аналітичними здібностями міг коректно ідентифікувати та узагальнювати таку інформацію, яку прості парсери часто не можуть розпізнати, якщо формат даних є нестандартним.

Одержаний набір даних із тисячами сполук, кожна з яких має численні анотовані властивості, є одним з найповніших ресурсів у галузі термоелектрики на сьогодні. Автори підкреслюють, що використання LLM дозволило подолати багато обмежень ручного або напівавтоматичного збору інформації. Наприклад, модель могла адаптуватись до різних форм подання даних (різні одиниці, порядок слів, формулювання), що було проблемою для ручних підходів.

Після автоматичного вилучення даних, автори запровадили скрипти для стандартизації одиниць вимірювання. Крім того, було здійснено автоматизоване перехресне порівняння «витягнутої» інформації з вихідним текстом, що підвищило надійність бази даних. Незважаючи на те, що точні показники точності ще не були оприлюднені, на основі попереднього досвіду очікується висока точність.

Важливим покращенням стала структурованість: кожен запис бази має чітке маркування контексту (склад, структура, властивості, температура тощо). Завдяки цьому, база стала особливо корисною для навчання машинних моделей та аналізу зв'язків у матеріалознавстві. Наприклад, тепер можна швидко знайти всі матеріали з кубічною структурою типу алмазу та коефіцієнтом Зеєбека понад 200 μV/K при 300 K, що раніше потребувало значних зусиль із пошуку в літературі. Авторами наголошується, що залишаються виклики щодо повністю автоматизованого збору даних, зокрема розмежування даних кількох матеріалів у межах одного тексту та визначення того, чи отримані дані експериментально, чи розраховані теоретично. Ці питання є предметом подальших досліджень та покращень.

Висновки

1. Існують рішення для формування достатньої бази властивостей термоелектричних матеріалів із наукової літератури на основі поєднання автоматичних інструментів із ШІ та верифікації за допомогою більш спеціалізованих інструментів чи людини.
2. Потенційні напрямки для подальшого удосконалення технології: обробка табличних та графічних елементів наукових статей, робота із іншомовними даними, оскільки переважна кількість LLM моделей було навчено саме на англомовній літературі, подальша оптимізації методів пошуку задля скорочення затребуваних обчислювальних потужностей.
3. Перспективним є розробка ШІ агента, здатного безперервно аналізувати нові статті у термоелектричному матеріалознавстві, що дозволить науковцям отримати миттєвий доступ до актуального джерела інформації.

Інформація про авторів

Короп М.М. – Аспірант.
Прибила А.В. – Кандидат фіз.-мат. наук.

Література

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

Надійшла до редакції 27.01.2025

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

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.

Submitted: 27.01.2025