


Research article

DOI: <https://doi.org/10.18721/JCSTCS.18305>

UDC 004.89



METHOD FOR AUTOMATED ENRICHMENT OF A KNOWLEDGE BASE ON GLASS COMPOSITIONS AND PROPERTIES BASED ON DATA FROM SCIENTIFIC PUBLICATIONS

E.A. Pavlov , *P.D. Drobintsev*, *V.A. Klinkov*,
A.V. Semencha, *I.G. Chernorutskiy*

Peter the Great St. Petersburg Polytechnic University,
St. Petersburg, Russian Federation

 pavlov_ea@spbstu.ru

Abstract. Automating the extraction of glass composition and property data from scientific literature is critically important for accelerating the development of new material. This work presents a method integrating: 1) the collection of full-text articles using the Elsevier Research Products APIs, 2) text preprocessing, 3) context-dependent extraction of structured data using a large language model (LLM) and a domain-specific prompt, 4) enrichment of a knowledge base on glasses. The key achievement is the development of a prompt that yields an F1-score of 0.99 for extracting chemical compositions, their properties and correctly establishing relationships between them on a sample of 50 articles. The proposed method significantly simplifies the automatic creation and continuous updating of knowledge bases on glasses, thereby eliminating the traditional reliance on manually curated, potentially outdated resources and providing a robust, data-driven foundation for the efficient designing of glasses with target properties using machine learning.

Keywords: data extraction, natural language processing, LLM, prompt engineering, knowledge base, glass, glass properties

Citation: Pavlov E.A., Drobintsev P.D., Klinkov V.A. et al. Method for automated enrichment of a knowledge base on glass compositions and properties based on data from scientific publications. Computing, Telecommunications and Control, 2025, Vol. 18, No. 3, Pp. 58–67. DOI: 10.18721/JCSTCS.18305


Научная статья

DOI: <https://doi.org/10.18721/JCSTCS.18305>

УДК 004.89



МЕТОД АВТОМАТИЗИРОВАННОГО ПОПОЛНЕНИЯ БАЗЫ ЗНАНИЙ О СОСТАВАХ И СВОЙСТВАХ СТЕКОЛ НА ОСНОВЕ ДАННЫХ ИЗ НАУЧНЫХ ПУБЛИКАЦИЙ

Е.А. Павлов , П.Д. Дробинцев, В.А. Клинков,
А.В. Семенча, И.Г. Черноруцкий

Санкт-Петербургский политехнический университет Петра Великого,
Санкт-Петербург, Российская Федерация

 pavlov_ea@spbstu.ru

Аннотация. Автоматизация извлечения данных о составах и свойствах стекол из научной литературы критически важна для ускорения разработки новых материалов. В работе представлен метод, интегрирующий: 1) сбор полнотекстовых статей с помощью Elsevier Research Products APIs, 2) предобработку текста, 3) контекстно-зависимое извлечение структурированных данных с помощью большой языковой модели (LLM) и доменно-специфичного промпта, 4) пополнение базы знаний о стеклах. Ключевым достижением стала разработка промпта, обеспечивающего точность $F1=0,99$ для извлечения химических составов и их свойств, а также корректного установления связей между ними на выборке из 50 статей. Предлагаемый метод значительно упрощает автоматическое создание и непрерывное обновление баз знаний о стекле, тем самым устраняя традиционную зависимость от вручную отобранных, потенциально устаревших ресурсов и обеспечивая надежную, управляемую данными основу для эффективного проектирования стекол с заданными свойствами с помощью машинного обучения.

Ключевые слова: извлечение данных, обработка естественного языка, большая языковая модель, промпт-инжиниринг, база знаний, стекло, свойства стекла

Для цитирования: Pavlov E.A., Drobintsev P.D., Klinkov V.A. et al. Method for automated enrichment of a knowledge base on glass compositions and properties based on data from scientific publications // Computing, Telecommunications and Control. 2025. Т. 18, № 3. С. 58–67. DOI: 10.18721/JCSTCS.18305

Introduction

Contemporary materials science research, particularly in designing functional glasses for optics, electronics and energy, faces an exponential growth of published data. Critical data on chemical compositions and physicochemical properties needed for predicting and designing new glasses remain “locked” in unstructured text of scientific publications [1]. Manual data collection and sorting requires a lot of labor-intensive: an expert may require up to one hour to thoroughly analyze a single article. This approach not only slows progress but also makes real-time analysis of thousands of publications impossible. Extracting data for multicomponent glass systems is especially challenging, where compositions are described in diverse, often non-standard formats (linear combinations, variable-based formulas, at.%/mol.% etc.) and property values heavily depend on synthesis and measurement methods. This specificity hinders traditional automated extraction methods. Large-scale, well-characterized data is fundamental for ensuring the accuracy and reliability of material inferences [2]. For example, the SciGlass database [3], long considered the gold standard for glass properties, is no longer updated, rendering it unsuitable for analyzing modern research. The lack of up-to-date structured glass composition and property data impedes the development of new materials with tailored properties. Given the rapidly expanding volume of scientific literature on glasses, automated extraction of composition and property data has become increasingly vital.

Breakthroughs in natural language processing (NLP), driven by the transformer architecture and attention mechanisms [4], have opened new avenues for textual data analysis [5–10]. Large language models (LLMs) like GPT [11], BERT [12] and Falcon [13] demonstrate unique capabilities in identifying semantic relationships and extracting hidden patterns from weakly structured texts. LLM-based methods often surpass classical NLP approaches in tasks requiring deep contextual understanding. Traditional entity extraction pipelines (e.g., OSCAR [14], ChemicalTagger [15], ChemDataExtractor [16]) are effective for text mining in chemistry and materials science but are limited by rigid templates and poor adaptation to semantic nuances. A key advantage of LLMs – particularly valuable for domains with complex data structures – is flexible extraction adaptation via prompt engineering techniques, enabling task-specific tailoring without resource-intensive fine-tuning [17, 18].

Current research confirms LLM effectiveness for scientific publication data extraction [19–21]. For example, in [19], BERT is refined on battery publications, combining question-answering with classical NLP to extract battery components (cathode, anode, electrolyte) and link them to characteristics like capacity and cycle stability, improving specialized database population accuracy by 18%. In [20], a ChemPrompt Engineering approach is proposed using ChatGPT to automate extraction of metal-organic framework (MOF) synthesis conditions. In [21], a pipeline for extracting reticular material (MOF, COF) synthesis parameters from Elsevier API-sourced PDFs using language models is implemented. While existing studies confirm LLM efficacy for chemistry and materials science data extraction, current solutions either require domain-specific model fine-tuning or lack adaptation to key features of glass system descriptions, such as diverse composition formats, mandatory unit handling and measurement method dependencies.

This article describes an automated knowledge base enrichment method combining an LLM guided by a specially designed domain-specific prompt without additional training or fine-tuning. The method enables intelligent digital platforms capable of near-real-time analysis of thousands of publications. This is particularly relevant for AI-driven design of glasses with target properties, where data quality and volume directly impact accuracy.

Method

The ultimate goal of this work is to create a continuously updated knowledge base on glasses. The target data schema is designed to store not only individual facts but also the complex relationships critical for analysis and predictive modeling. Its core consists of the following entities: detailed records of chemical compositions, specifying the concentration type (at.%, wt.%, mol.%), material properties, each accompanied by a value, unit of measurement and, crucially, the measurement method, ensuring correct comparison of data from different sources and enriched publication metadata (authors, journal, year). To transform the unstructured text of scientific articles into such a relational structure, an end-to-end pipeline was developed. Its key task is not merely to extract mentions of substances and numbers, but to correctly interpret them in context, identify relationships between them and package them into a strict, formalized format (JSON) suitable for automatic analysis using the following pipeline.

Pipeline

The proposed method for automated glass property knowledge base formation is implemented as follows (Fig. 1):

1. Article Retrieval: Full-text articles in XML format are downloaded from ScienceDirect via Elsevier Research Products APIs.
2. Preprocessing: Article text is processed. Only main text and metadata are extracted, with irrelevant elements removed (figure references, bibliographies, etc.).
3. LLM Text Processing: Preprocessed text is fed into an LLM with a prompt generating structured JSON of a specified format.
4. Postprocessing: The LLM response is processed, and extracted information is saved to a database.

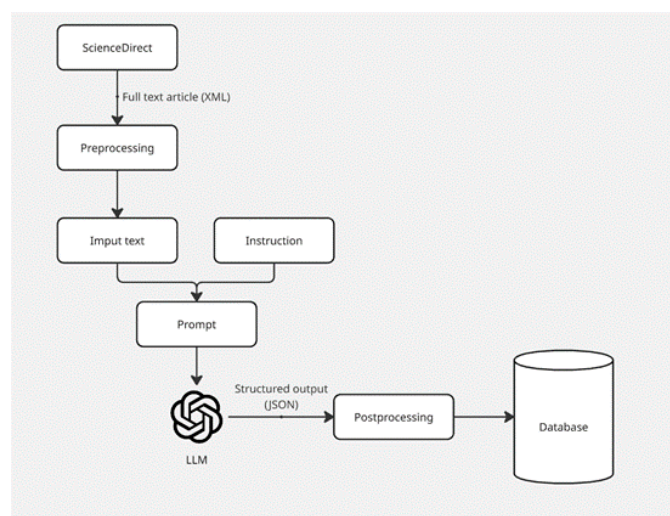


Fig. 1. Automated knowledge base formation pipeline

Article extraction via Elsevier API

The Elsevier API is utilized to access full-text scientific publications from the ScienceDirect library to extract information on glass compositions and properties. When developing automated scientific article processing systems using language models, the selection of an input data format that ensures accurate interpretation of information, particularly tables, is crucial. Although several formats (JSON, PDF, XML, TEXT) are available via the Elsevier API, the XML format was selected in this study. This choice was driven by the issue of incorrect recognition of missing values in tables encountered when using other data formats, which adversely affected information extraction quality.

To collect relevant publications, a search query was executed via the Elsevier API, incorporating the following keyword combination: “glass”, “chalcogenide glass”, “As-Se”, “Tg”, “density”. To minimize noise and enhance the relevance of input data fed into the language model, article text preprocessing was implemented. This preprocessing involves removing information unrelated to material compositions and properties. Specifically, only tables and the main body text are extracted from the source XML. This approach is necessary to ensure the model focuses on the core article content – where methods, results and discussion are presented – and to reduce the volume of input data processed by the model. The output is a cleaned text containing exclusively data relevant for analysis.

Information extraction

The Qwen 3 LLM was selected for automating structured glass composition/property extraction. Qwen 3 employs a hybrid approach with two modes:

- **Thinking Mode:** The network spends time reasoning step-by-step before final output.

This mode is essential for decomposing intricate material descriptions, resolving ambiguities in terminology or units, and inferring implicit relationships between composition and properties – common challenges in scientific text extraction.

- **Non-Thinking Mode:** Delivers rapid “near-instant” responses (suited for simple queries prioritizing speed).

Qwen 3 offers revolutionary advances in logical reasoning, instruction following, agent capabilities and multilingual support. The explicit reasoning chain in Thinking Mode significantly enhances the reliability and accuracy of extracted data from dense, research papers. It is open-source and free to use.

Prompt engineering

Prompt engineering is crucial for extracting glass composition/property data from publications. It precisely formulates queries for LLMs, minimizing hallucination risks [22] and improving accuracy. For

complex data structures, prompts must explicitly instruct the LLM to output extracted information in a strict format. JSON is optimal due to its ubiquity in LLM training data, ensuring correct structuring and simplifying downstream processing. Thus, effective prompt design optimizes extraction and reduces error risks.

An iterative prompt development approach was applied. The prompt underwent several refinement cycles using a validation set of 10 articles, which were excluded from the final test sample. The key stages of this refinement process are described below. The initial prompt defined for the LLM requested extraction of all glass compositions and properties in JSON format.

Initial prompt:

Analyze the provided text and extract information about the compositions and their properties. Present the result in JSON format, grouping the properties by each composition. If the text mentions multiple compositions, create a separate object for each one. For each composition, specify its name and a list of properties, where each property contains a name and a value (if specified). If the property value is not specified, leave the field empty or use null.

This prompt formulation revealed two key limitations:

1. Inconsistent JSON formatting. The model returned JSON objects with varying structures, complicating subsequent processing.
2. Incorrect handling of chemical formulas. The model failed to interpret complex composition expressions, such as $(100 - x)[70\text{GeO}_2 - 12\text{BaF}_2 - 10\text{Ga}_2\text{O}_3 - 3\text{BaO} - 3\text{La}_2\text{O}_3 - 2\text{Y}_2\text{O}_3] - x\text{Dy}_2\text{O}_3$, and to calculate elemental composition percentages.

To overcome these issues, the prompt was significantly enhanced. The core modifications included:

- Explicit instructions for composition parsing.

The LLM was explicitly tasked to identify different compositional formats (e.g., linear combinations, percentage-based notations) and to calculate elemental percentages using provided formulas. A critical addition was the requirement to determine and output the type of percentage (wt%, mol%, at%) for each composition, as this is essential for interpreting the data correctly.

...

Instructions:

1. Composition Identification

– Detect all compositional formats:

a) Linear combination: $(1 - x) \cdot [\text{Base}] + x \cdot [\text{Additive}]$ (e.g., $(1 - x) \cdot \text{Ge}_{20}\text{As}_{20}\text{Se}_{40} + x \cdot \text{Te}$)

b) Percentage-based notation: $\text{Ge}_{12}\text{As}_{24}\text{Se}_{64}$ or $\text{As}_{30}\text{Se}_{40}\text{Te}_{30}$

– Determine percentage type (wt/atomic/mol. %)

– For linear combinations:

* Calculate wt/atomic/mol. % using:

$\text{Element\%} = (1 - x) \cdot (\text{base\%}) + x \cdot (\text{additive\%})$

* Round to 2 decimal

– For percentage-based notation: Keep original wt/atomic/mol. percentages

...

- Strict JSON schema definition.

A comprehensive JSON example was embedded into the prompt to enforce a consistent output structure. The schema requires each composition entry to include the following keys:

- “type”: “calculated” for expressions requiring computation (e.g., linear combinations), “raw” for direct notations.
- “percentage_type”: The unit of the composition's concentration (“mol%”, “wt%”, “at%”).
- “formula”: The original compositional expression as found in the text.
- “x”: The value of the parameter x for parametrized compositions.
- “composition”: A key-value object where each key is a chemical element/compound and its value is the calculated or extracted percentage.

- “properties”: A key-value object where each key is a property name. Each property must be an object containing the keys “value” and “unit”.

...

JSON Structure

```

    "json
    {
      "compositions": {
        "GAST-1": {
          "type": "calculated",
          "percentage_type": "mol%",
          "formula": "(1 - 0.2)·Ge25As25Se50 + 0.2·Te",
          "x": 0.2,
          "composition": {"Ge": 20.0, "As": 20.0, "Se": 40.0, "Te": 20.0},
          "properties": {"density": {"value": 3.45, "unit": "g/cm3"},
            "Tg": {"value": 285, "unit": "°C"}}
        },

```

...

However, this prompt formulation also proved suboptimal. Further testing revealed the following issues:

- Hallucination on missing compositions.

When no chemical compositions were found in the article text, the LLM occasionally hallucinated by returning the example response provided in the prompt.

- Omission of duplicate properties.

The LLM skipped properties with identical names. For instance, if glass transition temperature was measured using multiple methods in an article, only one value was extracted. The same behavior occurred with properties like refractive index.

These issues were resolved by adding explicit instructions to the prompt:

- Return empty JSON if no compositions are detected.

...

4. JSON Structure

– If NO COMPOSITIONS detected: return {}

– For each detected composition:

```

    "json

```

...

- For multiple measurements of the same property, require the LLM to include the full property name and measurement method in its output.

...

– Include the measurement method from the text. If method is unspecified → "measurement_method": "Not specified".

– Map abbreviations to full names (e.g., "Eg" → "Optical Bandgap", "HV" → "Vickers Hardnes")

– Format properties as key-value pairs:

```

    "json
    "properties": {"property_abbreviation": {
      "property_full_name": "...",
      "value": ...,
      "unit": "...",
      "measurement_method": "..."}

```

...

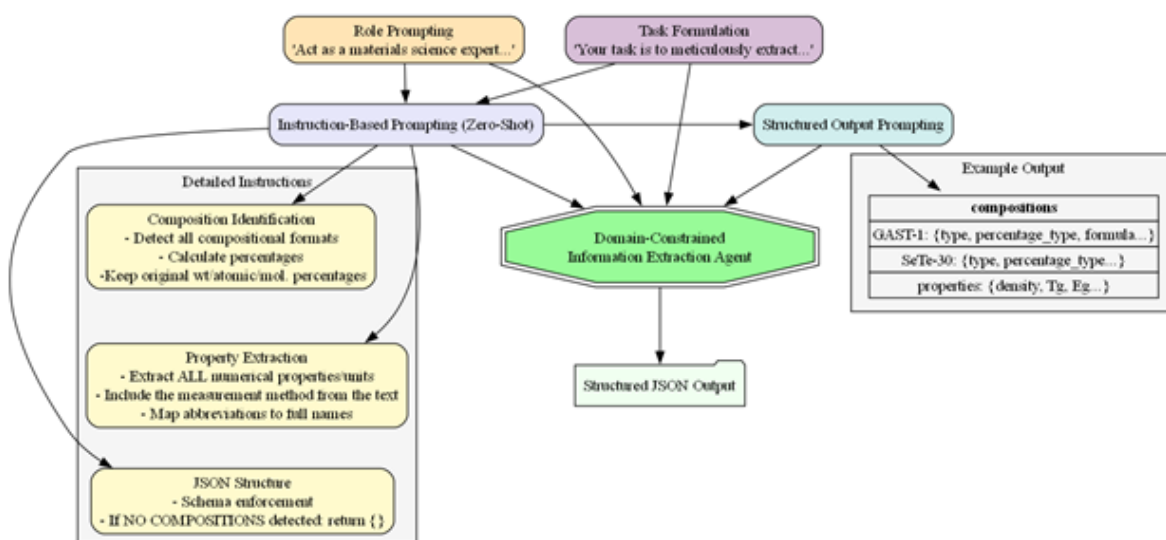


Fig. 2. Prompting architecture

Consequently, the final prompt for extracting glass compositions and properties was formulated and is available in the GitHub repository associated with this work¹.

The extraction of precise compositional and property data from unstructured glass science literature necessitates overcoming significant challenges, including variable notation formats, domain-specific terminology and specific requirements for outputs. To address these, a multi-faceted prompt engineering strategy was implemented, combining established techniques to optimize LLM performance.

Role prompting was employed to instill domain-specific expertise, achieved through the explicit directive: “Act as a materials science expert specializing in glasses”. A precise task formulation followed, stating: “Your task is to meticulously extract compositional data and property values from research papers”. This explicit framing focused the model’s processing exclusively on targeted data extraction, minimizing diversion toward ancillary content within source texts. Instruction-based prompting under a zero-shot paradigm provided detailed procedural rules without in-context examples. Structured output prompting enforced rigorous schema compliance through a predefined JSON structure. The synergistic integration of these techniques transformed the LLM into a domain-constrained information extraction agent, as visualized in Fig. 2. Role prompting established foundational expertise, task formulation defined the operational objective, instruction-based rules governed the technical execution, and structured output prompting ensured machine-processable structured output. This layered methodology enabled reliable, automated curation of materials data from complex research texts.

Postprocessing

The JSON response from the LLM is merged with article metadata extracted from the source XML text and undergoes further processing before storage in the knowledge base. This processing includes JSON parsing, data validation, linking of the LLM’s JSON response to article metadata and persistence of validated records to a relational database.

Core stored entities comprise:

1. Publication Metadata:

- Article title
- Author list
- Journal name
- Publication year

¹ <https://github.com/EvgenDI/automated-glass-data-extraction>

- Page range
2. Extracted Data:
- Chemical compositions
 - Properties (value, unit of measurement, measurement method, full descriptive name)

Results and discussion

To evaluate the quality of glass composition and property extraction from scientific publications, 50 articles were selected. A materials science expert manually extracted glass composition data, corresponding properties, and their relationships from these articles, creating a benchmark dataset containing 253 compositions and 1685 properties.

Testing employed a local Qwen-3-14B model with chain-of-thought reasoning enabled. The model was deployed on a system with 1× Tesla A100 80GB GPU, 8 CPU cores, and 243 GB RAM. Standard information extraction metrics were used for robust and standardized evaluation: precision, recall and F1-score.

- Precision reflects the proportion of relevant information among all extracted data. High precision indicates most extracted information is correct (few false positives).
- Recall indicates the proportion of available relevant information successfully extracted. High recall signifies minimal omission of relevant information (few false negatives).
- F1-score is the harmonic mean of precision and recall. This metric provides a balanced performance assessment by considering both extraction accuracy and completeness.

Results (Table 1) confirm the method’s high reliability:

- Composition identification precision: 100% (all extracted compositions/properties matched the benchmark)
- Only 1% of glass compositions were omitted.
- Correct relationships were established for 99% of accurately extracted compositions and properties (pair F1-score = 0.99).

This relational accuracy is critical for building trustworthy knowledge bases.

Table 1

Information extraction quality assessment

Metric	Compositions	Properties	Composition/Property
Precision	1	1	0.99
Recall	0.99	0.99	0.99
F1	0.99	0.99	0.99

The high F1-score underscores the potential for automating glass data extraction, reducing manual processing time. These results demonstrate that modern LLMs can significantly accelerate expert knowledge extraction while maintaining high reliability.

Conclusion

This study developed a method that integrates automated full-text article extraction from ScienceDirect with the application LLM and engineered domain-specific prompt. The developed method demonstrated high accuracy in extracting structured data on glass composition and properties from scientific publications. Testing on a separate set of 50 articles yielded an F1-score of 0.99, demonstrating its high efficacy. The capabilities of the method significantly facilitate the transition from manual data collection to automated knowledge extraction. This achievement paves the way for creating self-learning systems capable of accelerating the development of glasses with targeted properties.

REFERENCES

1. Krallinger M., Rabal O., Lourenço A., Oyarzabal J., Valencia A. Information retrieval and text mining technologies for chemistry. *Chemical Reviews*, 2017, Vol. 117, No. 12, Pp. 7673–7761. DOI: 10.1021/acs.chemrev.6b00851
2. Hill J., Mulholland G., Persson K., Seshadri R., Wolverton C., Meredig B. Materials science with large-scale data and informatics: Unlocking new opportunities. *MRS Bulletin*, 2016, Vol. 41, Pp. 399–409. DOI: 10.1557/mrs.2016.93
3. Mazurin O.V., Leko V.K., Streltsina M.V., Shvayko-Shvaykovskaya T.P. Sovremennoe sostoianie bazy dannykh informatsionnoi sistemy SciGlass v oblasti opticheskikh kharakteristik stekol [The current state of the SciGlass information system database in the field of optical characteristics of glasses]. *Nauchno-tekhnicheskii vestnik sankt-peterburgskogo gosudarstvennogo universiteta informatsionnykh tekhnologii, mekhaniki i optiki* [Scientific and Technical Bulletin of Information Technologies, Mechanics and Optics], 2004, No. 13, P. 253.
4. Foppiano L., Lambard G., Amagasa T., Ishii M. Mining experimental data from materials science literature with large language models: an evaluation study. *Science and Technology of Advanced Materials: Methods*, 2024, Vol. 4, No. 1, Art. no. 2356506. DOI: 10.1080/27660400.2024.2356506
5. Peng R., Liu K., Yang P., Yuan Z., Li S. Embedding-based retrieval with LLM for Effective Agriculture Information Extracting from Unstructured Data. *arXiv:2308.03107*, 2023. DOI: 10.48550/arXiv.2308.03107
6. Patiny L., Godin G. Automatic extraction of FAIR data from publications using LLM. *ChemRxiv*, 2023. DOI: 10.26434/chemrxiv-2023-05v1b-v2
7. Biswas A., Talukdar W. Robustness of structured data extraction from in-plane rotated documents using multi-modal Large Language Models (LLM). *arXiv:2406.10295*, 2024. DOI: 10.48550/arXiv.2406.10295
8. Birhane A., Kasirzadeh A., Leslie D., Wachter S. Science in the age of large language models. *Nature Reviews: Physics*, 2023, Vol. 5, Pp. 277–280. DOI: 10.1038/s42254-023-00581-4
9. Manjotho A.A., Tewolde T.T., Duma R.A., Niu Z. LLM-guided fuzzy kinematic modeling for resolving kinematic uncertainties and linguistic ambiguities in text-to-motion generation. *Expert Systems with Applications*, 2025, Vol. 279, Art. no. 127283. DOI: 10.1016/j.eswa.2025.127283
10. Chen X., Huang X., Gao Q., Huang L., Liu G. Enhancing text-centric fake news detection via external knowledge distillation from LLMs. *Neural Networks*, 2025, Vol. 187, Ar. no. 107377. DOI: 10.1016/j.neunet.2025.107377
11. Achaim J., Adler S., Agarwal S. et al. GPT-4 Technical Report. *arXiv:2303.08774*, 2023. DOI: 10.48550/arXiv.2303.08774
12. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv:1810.04805*, 2018. DOI: 10.48550/arXiv.1810.04805
13. Ehtesam A., Hamza A., Abdulaziz Al. et al. The Falcon Series of Open Language Models. *arXiv:2311.16867*, 2023. DOI: 10.48550/arXiv.2311.16867
14. Jessop D.M., Adams S.E., Willighagen E.L., Hawizy L., Murray-Rust P. OSCAR4: a flexible architecture for chemical text-mining. *Journal of Cheminformatics*, 2011, Vol. 3., Art no. 41. DOI: 10.1186/1758-2946-3-41
15. Hawizy L., Jessop D.M., Adams N., Murray-Rust P. ChemicalTagger: A tool for semantic text-mining in chemistry. *Journal of Cheminformatics*, 2011, Vol. 3, Art. no. 17. DOI: 10.1186/1758-2946-3-17
16. Mavračić J., Court C.J., Isazawa T., Elliott S.R., Cole J.M. ChemDataExtractor 2.0: Autopopulated ontologies for materials science. *Journal of Chemical Information and Modeling*, 2021, Vol. 61, No. 9, pp. 4280–4289. DOI: 10.1021/acs.jcim.1c00446
17. Chen B., Zhang Z., Langrené N., Zhu S. Unleashing the potential of prompt engineering for large language models. *Patterns*, 2025, Vol. 6, No. 6, Art. no. 101260. DOI: 10.1016/j.patter.2025.101260
18. Knoth N., Tolzin A., Janson A., Leimeister J.M. AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence*, 2024, Vol. 6, Art. no. 100225. DOI: 10.1016/j.caeai.2024.100225

19. **Huang S., Cole J.M.** BatteryBERT: A pretrained language model for battery database enhancement. *Journal of Chemical Information and Modeling*, 2022, Vol. 62, No. 24, Pp. 6365–6377. DOI: 10.1021/acs.jcim.2c00035
20. **Zheng Z., Zhang O., Borgs C., Chayes J.T., Yaghi O.M.** ChatGPT Chemistry assistant for text mining and the prediction of MOF synthesis. *Journal of the American Chemical Society*, 2023, Vol. 145, No. 32, Pp. 18048–18062. DOI: 10.1021/jacs.3c05819
21. **Da Silva V.T., Rademaker A., Lioni K., Giro R., Lima G., Fiorini S., Archanjo M., Carvalho B.W., Neumann R., Souza A., Souza J.P., de Valnizio G., Paz C.N., Cerqueira R., Steiner M.** Automated, LLM enabled extraction of synthesis details for reticular materials from scientific literature. *arXiv.2411.03484*, 2024. DOI: 10.48550/arXiv.2411.03484
22. **Ji Z., Lee N., Frieske R., Yu T., Su D., Xu Y., Ishii E., Bang Y.J., Madotto A., Fung P.** Survey of hallucination in natural language generation. *arXiv.2202.03629*, 2022. DOI: 10.48550/arXiv.2202.03629

INFORMATION ABOUT AUTHORS / СВЕДЕНИЯ ОБ АВТОРАХ

Evgeniy A. Pavlov

Павлов Евгений Алексеевич

E-mail: pavlov_ea@spbstu.ru

ORCID: <https://orcid.org/0000-0002-7437-6153>

Pavel D. Drobintsev

Дробинцев Павел Дмитриевич

E-mail: drob@ics2.ecd.spbstu.ru

Victor A. Klinkov

Клинков Виктор Артемович

E-mail: klinkovvictor@yandex.ru

Alexander V. Semench

Семенча Александр Вячеславович

E-mail: asemencha@spbstu.ru

Igor G. Chernorutskiy

Черноруцкий Игорь Георгиевич

E-mail: igcher1946@mail.ru

Submitted: 01.07.2025; Approved: 12.09.2025; Accepted: 19.09.2025.

Поступила: 01.07.2025; Одобрена: 12.09.2025; Принята: 19.09.2025.