UNIVERSITY OF TURKU

The Use of Large Language Models for Information Extraction

Department of Mechanical and Materials Engineering

Bachelor's thesis

Author:
Tiia Heino

13.6.2025

Turku

Bachelor's thesis

Subject: Materials engineering
Author: Tiia Heino
Title: The Use of Large Language Models for Information Extraction
Supervisor: Tomasz Galica
Number of pages: 21 pages
Date: 1.6.2025

Large Language Models (LLM) are Artificial Intelligence (AI) models that can be fine-tuned for specific tasks. They can be used for Information Extraction (IE), a domain of Natural Language Processing (NLP), that aims to extract information from unstructured text and convert it into structured, machine-readable data. Using LLMs for IE allows for more accurate and less time-consuming extraction. For materials science using LLMs for IE enables a more automated extraction that can be used with Materials Science Databases, that store diverse data on different materials. With these tools materials science research can go towards a new era with data-driven studies. Using LLMs does have its challenges with hallucinations, environmental costs and transparency issues.

Suuret kielimallit ovat tekoälymalleja, joita voidaan hienosäätää tiettyjä tehtäviä varten. Niitä voidaan käyttää tiedonlouhintaan, joka on luonnollisen kielen prosessoinnin osa-alue, jonka tavoitteena on poimia tietoa jäsentymättömästä tekstistä ja muuntaa se jäsennellyksi, koneella luettavaksi dataksi. Mallien käyttö tiedonlouhinnassa mahdollistaa tarkemman ja vähemmän aikaa vievän louhinnan. Materiaalitieteessä mallien käyttö tiedonlouhintaan mahdollistaa automatisoidumman louhinnan, jota voidaan käyttää Materiaalitieteellisten Tietopankkien kanssa, jotka sisältävät monipuolista tietoa eri materiaaleista. Näiden työkalujen avulla materiaalitieteellinen tutkimus voi siirtyä uuteen datapainoitteiseen aikakauteen. Suurten kielimallien käyttöön liittyy kuitenkin haasteita, joita ovat hallusinaatiot, ympäristökustannukset ja haasteet läpinäkyvyydessä.

Key words: Large Language Models, Information Extraction, Artificial Intelligence, materials science.

Table of contents

1 Introduction

The use of Artificial Intelligence (AI) has increased significantly over the last years both by private users as well as companies and researchers. There are different types of AI models, however on this thesis we'll be focusing on Large Language Models. AI is technology that enables computers and machines to mimic human abilities such as learning, understanding, problem solving, decision making, creativity and autonomous behaviour. Applications and devices powered with specific types of AI models can interpret and respond to human language. [1]

Machine learning (ML) is a subfield of AI that allows machines to learn from data without being explicitly programmed. ML models are used e.g. in image generation and natural language processing. Applications of ML in different industries range widely from healthcare to finance to science and more. ML brings the possibility to fundamentally change many industries by automating complex processes and making intelligent predictions and decisions. [2] In Figure 1 the timeline of development for the field can be seen with AI being created in the 1950's and it being followed by invention of ML, Deep Learning (DL), and Generative AI (Gen AI). [1]

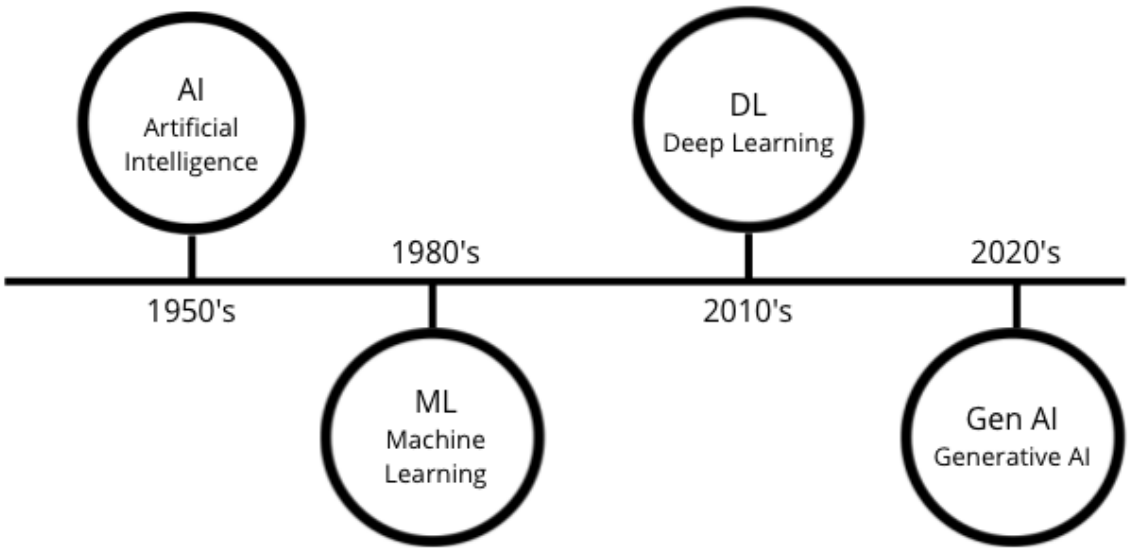

Figure 1. Timeline of how the field has developed from AI to Gen AI.

Data, and especially training data, is an integral part for AI and ML models and systems. It is a critical input that allows the models and systems to learn and improve in time. Without training data, the

models wouldn't be able to make predictions or extract needed information. With data being this big of a part for the models, the quality of it needs to be excellent to ensure a better result for the models. The first step in ML is collecting relevant data which can be from e.g. databases, sensors or the Internet. When the data has been collected, it is then pre-processed to ensure its quality and suitability for analysis. After pre-processing, an ML model is trained on pre-processed data, to make predictions or decisions. During the training, the model adjusts its learnable parameters. After the training, the model needs to be evaluated to assess its performance and see if it meets the required and desired criteria. To evaluate these models, we can use different approaches depending on the model and what is expected of it. Usually model predictions are evaluated with validation dataset. If the model is deemed to be successful, it can be deployed for real-world applications for ML. [1], [2], [3]

1.1 Large Language Models

Large Language Models (LLMs) are AI models that learn the syntax and semantics of natural language from massive text corpora [4]. LLMs are used e.g. in search engines, voice assistants and coding assistance tools [5]. LLMs must be fed a vast amount of textual data, called corpus, at their training phase to learn the complexities of natural language, so they can comprehend and produce natural language as, well as encode human knowledge into their weight parameters [4].

The idea of LLMs can be traced back to the early days of Natural Language Processing (NLP) and the understanding of semantics. NLP is an array of computational techniques for analysing and representing naturally occurring texts. Many applications require NLP, e.g. email filters, smart assistants like Apple's Siri and Amazon's Alexa and online translators. This brought forward a need for a model that could comprehend text input and generate output that is indistinguishable from human language. This need set in motion development, that eventually led to the invention of LLMs. [6], [7]

The evolution of LLMs began with studying how words connect within a language and how they convey meanings. Statistical Language Models (SLMs) were invented in the 1990's as mathematical models that used probability to predict and analyse natural language. The main point of SLMs is predicting how likely a sequence of words is, based on all the words which appeared before in the given sentence. Following the SLMs, Neural Language Models (NLMs) were invented to handle longer sequences and mitigate the limitations of SLMs. NLMs leverage neural networks to predict the probabilities of subsequent words in a sentence. With the introduction of pre-training, NLMs evolved into Pre-Trained Language Models (PLMs). PLMs undergo training with a vast amount of unlabelled text, which enables them to grasp fundamental language structures. To optimize the models' performance, the PLMs can be fine-tuned with a smaller dataset. The need for a larger scale model led to LLMs being invented. LLMs are trained with text corpora that has at least tens of billions of

parameters. The significant difference between PLMs and LLMs is the parameter count and training data volume that is over ten times bigger for LLMs. Generative LLMs are currently state-of-the-art, which is why the focus is on them for this thesis. Figure 2 provides a visual for this development of LMs. [7], [8]

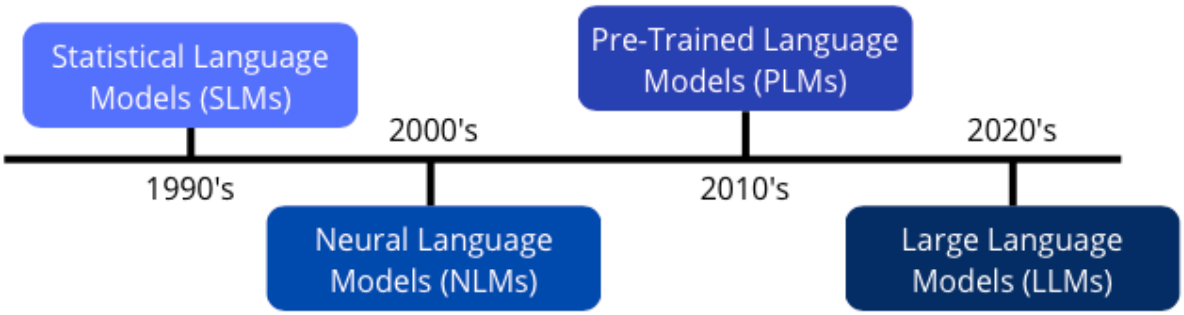

Figure 2. Timeline of the development of Language Models.

The core task of LLMs, text generation, is in an essence prediction of sequence of words (tokens). This prediction happens one step at a time based on the context and sequence of preceding tokens. This prediction process has five steps: inputting context, probability calculation, token selection, updating context and repeating. In the first step, the model receives an initial sequence of tokens that can be e.g. a prompt provided for it or text it has generated so far. Moving to the second step, the model analyses the patterns of the context. Then it calculates the probability for every possible token in its vocabulary. Next token can be selected with one of few methods. The model can use the simplest strategy, which means the model just picks the token with the highest calculated probability. More complex strategies can involve sampling e.g. the top ten choices with the highest probability for the next token to introduce variety. Next the selected token is added to the sequence. This process repeats itself until the model reaches a stopping condition or it fulfils a specified length given on the prompt. LLMs learn the token co-occurrence and probabilities during the training. [9]

The LLMs training combines two fundamental processes: pre-training and fine-tuning. This method of training ensures that LLMs are flexible and can be anything from general-purpose to very specialized models. Pre-training of LLMs uses massive unlabelled corpora to learn dependencies between words and sentences and this phase results in a foundation model. In fine-tuning the models are fed with labelled corpus that is allowing the model to adapt and improve its performance on specific tasks. Fine-tuning allows the models to generate appropriate answers on the specific themes that are in the labelled corpus. The process of training for LLMs has been visualized in Figure 3. [10]

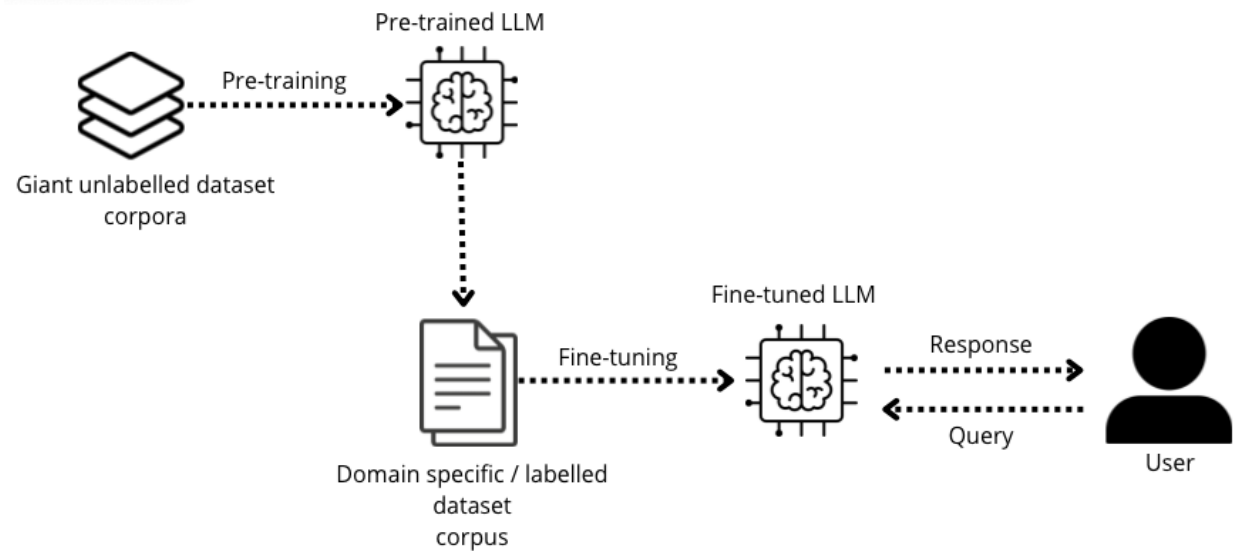

Figure 3. Visualization on how LLMs are trained all the way to fine-tuned LLMs.

LLMs can also be refined for conversational uses. Fine-tuning the models for this involves two essential features: prediction of next words and word differentiation. The first feature enables LLMs to predict the next word in a sequence using knowledge acquired during pre-training. In contrast, word differentiation allows LLMs to distinguish among multiple possible word choices through fine-tuning, refining their predictions to suit specific tasks or domains. Due to the complicated nature of human language, additional capabilities are required for the LLMs to converse with ease with humans. Typically, this is achieved by using reinforcement learning for the models. Reinforcement learning allows LLMs to acquire new skills through iterative learning its interactions. [10]

LLMs can be categorized based on how they are accessed: Open Source LLMs and Closed Source LLMs. Open Source LLMs allow anyone to access their code and algorithm, and they are built on the principalities of transparency and collaboration. In some cases, even the data that is used to train the LLMs, is available for anyone to access. They drive the idea that AI should be an accessible tool to everyone, not just those that have the resources to pay for it. Closed Source LLMs are developed by companies, e.g. Google and OpenAI, that keep their models' working methods private, instead offering them as paid services. Some main differences between these two different types of LLMs can be seen in Table 1. [11]

Table 1. Comparison of Open Source LLMs and Closed Source LLMs

FEATURE	OPEN SOURCE LLMS	CLOSED SOURCE LLMS
ACCESSIBILITY	freely available and accessible to all users	restricted access, often requires a payment

INNOVATION	community-driven, fast-paced development	centralized, controlled development processes
SECURITY	transparent, vulnerable for misuse	lacks external auditing and transparency, more secure from misuse
SUPPORT	community forums and documentation	professional support services and resources

1.2 Data extraction

Data extraction is the process of gathering and moving data from multiple sources to a single destination to be stored. These sources can range from publications to databases to spreadsheets etc. The extracted data can come in different forms, be unorganized or unstructured. The main point in data extraction is to bring together the fragmented information in a centralized location, e.g. on-site, cloud-based, or a location that is a combination of both. [12]

Data extraction can be divided into three main types, full extraction, incremental extraction and extraction of unstructured data. Full extraction is a straightforward method that retrieves all available data from the source directly. This ensures the data is correct and complete but can be resource-intensive and time-consuming. Incremental extraction captures only the changes in the data that have occurred since the last extraction. As a method it's more efficient than full extraction as it reduces the amount of data transferred and saves on data processing time. Incremental extraction can be implemented in two ways, either in batch or stream approach. Batch approach captures changes in chunks at defined intervals. Stream approach allows for real-time updates as it captures changes almost immediately as they happen. Extraction of unstructured data is more complex than the above-mentioned methods as there is a lack of standard structures and formats. Unstructured data sources are e.g. emails, PDF files and web pages. [12]

For materials science, especially the extraction of unstructured data is relevant. Full and incremental extractions are used for databases and similar sources; however, a vast majority of materials science data is stored in publications that are unstructured data. Using extraction of unstructured data allows the extraction from e.g. PDFs or web pages, the usual ways publications are available. Both full and incremental extraction can be used as well. Incremental extraction could be very useful for extraction of data that is often updated e.g. when extracting data from a database that has been used for the same project or research already.

LLMs enabled easier extraction of information from unstructured textual data. Using prompt-engineering, a method to modify LLM instructions (prompts), we can optimize information extraction done with these models. This combination ultimately allows for a more accurate and less resource consuming data extraction. [13]

2 Data In Materials Science

Materials science is the study of characteristics and applications of materials that combines chemistry, physics and engineering research. In the traditional approach materials discovery and research, is carried out through conventional experimental, theoretical or computational research. This is usually done in a trial-and-error fashion, which takes up a lot of time and can make researching expensive. The longer it takes to get the results the more resources and testing it takes which then leads to more costs. [14]

Data-driven studies is the practice of using advanced analysis tools combined with already existing data, guiding the process of the research. By employing both computational and experimental methods for data generating as well as employing ML techniques, data-driven studies have led materials science into a new era. It differs from the traditional approach by the volume of data and the more automated way of conducting experiments. [14], [15]

Table 2. Comparison of Traditional approach and Data-driven studies

ASPECT	TRADITIONAL APPROACH	DATA-DRIVEN STUDIES
BASIS	fundamental physics	data, AI
SPEED	slow (months/years)	fast (hours/days)
COST	expensive	less expensive (not cheap)
SCALABILITY	limited by experiments	can handle big data

As seen in Table 2, the volume of data and the way information is extracted makes data-driven studies faster than the traditional approach. Both approaches are expensive, but as data-driven approach is more cost and time efficient by utilizing already existing knowledge, it's the cheaper option of the two. Both approaches are vital, e.g. the traditional approach can validate hypotheses and data-driven approach can speed up discovery pipelines, and they can be used together to strengthen each other.

2.1 Materials Science Databases

Materials Science Databases play a critical role in accelerating research by providing structured access to experimental, computational and literature-based data. They enable scientists and engineers to access, store and analyze vast amount of data from various sources e.g. simulations, experiments and literature. The databases have diverse information on different materials that can include details on their structures, properties and reaction mechanisms. In addition to storing data, databases can support interdisciplinary collaboration across diverse fields and between their experts, e.g. experimental

researchers, theoretical scientists and programmers, allowing for more efficient and effective problem-solving. [16]

Materials databases can be categorized into groups by their content, purpose or data sources. Some of the commonly used categories are experimental databases, computational databases, hybrid databases and ML-oriented databases. Experimental databases store data obtained from experiments, and the content can be e.g. material's crystal structure, mechanical properties or thermal conductivity. Computational databases on the other hand contain data that has been generated through computational simulations. They can include data on material's mechanical properties or thermal behaviour. As is in the name, hybrid databases are a combination of both experimental and computational data, providing a broader dataset. ML-oriented databases are designed for use in machine learning applications, and they provide curated datasets. In Table 3 different databases are presented with their content and to what category they belong to. [16], [17]

Table 3. A list of some databases and what they contain and what category they belong to.

DATABASE	CONTENT	CATEGORY
THE AUTOMATIC FLOW FOR MATERIALS DISCOVERY (AFLOW)	over 3 million high-throughput calculations	computational
THE MATERIALS PROJECT (MP)	calculated data for over 140 000 materials	computational
MATNAVI	data on polymers, metals and inorganic crystals	hybrid
NOMAD	files for more than 50 different atomistic codes, more than 100 million total-energy calculations	ML-oriented

With these databases LLMs can be used to automize data retrieval and summarization, saving time and resources. The models can be used with Retrieval Augmented Generation (RAG), which is a method that enhances LLMs abilities by integrating information from external knowledge sources. RAG was developed to improve response accuracy and information extraction from text, which especially with materials science adds value to using LLMs as it makes the models and their outputs more stable. For example, LLamP Framework is a multimodal framework that combines RAG with multimodal agents, that has been invented especially for materials science researching [4]. The multimodal agents access sources like The Materials Project. This framework greatly improves material property predictions and allows for generation of realistic 3D models that LLMs alone struggle with. To be able to retrieve information from databases, LLMs need to use Information Extraction. [4]

3 Information Extraction

Information Extraction (IE) is a domain of NLP that aims to extract information from unstructured text systematically and converting it into structured, machine-readable data. To achieve this conversion IE identifies and categorizes key elements e.g. entities, events and numerical values. The field has evolved from rule-based systems to advanced Machine Learning and Deep Learning techniques. As LLMs have a proficiency in context understanding and an ability to organize information, they have a strong adaptability in IE. [18]

The first IE systems were mostly rule-based, which meant they were efficient in specific context as they relied on predefined rules to extract information. This meant they were not great with scalability nor adaptability. Advances in ML brought supervised learning models (SLMs) that offered stronger and more adaptable solutions for IE. But even with SLMs, the dependency on large, annotated datasets and problems on capturing deeper linguistic contexts remained as challenges. With deep learning, new models were introduced to IE, that were able to enhance the accuracy and efficiency of IE-systems by capturing complex patterns and dependencies in text. Deep learning models reduced the reliance on hand-crafted features. From advanced deep learning models, IE moved to using LLMs. Performance of IE is usually evaluated in terms of precision, recall and F-score. Definition of percision is the percentage of retrieved results that are correct. Recall is the percentage of correct results that are retrieved. F-score is harmonic mean of precision and recall. [19] [18]

Using LLMs for IE enables automated retrieval of structured data from unstructured or semi-structured text. To achieve the most efficient result from IE with LLMs, the models should extract the most accurate data with the least amount of errors. LLMs for IE can be fine-tuned for domain-specific tasks, combined with prompt-engineering and used with hybrid pipelines. Fine-tuning the models with annotated datasets makes them domain-aware or specialized in a given task. Using LLMs with prompt-engineering, improves the obtained results. In the study of Hu et al. the F1 score of LLM GPT-4 was 0.804 without prompt-engineering and improved 5.7 % to 0.861 with prompt-engineering [20]. Hybrid pipelines combine LLMs with traditional NLP to reduce hallucinations and improve efficiency. By combining these two we get e.g. name entity recognition (NER) models like MaterialsBERT, that has been invented specifically for materials science data tasks. [13], [20], [21]

3.1 Data mining

Data mining is the central part of the method of knowledge detection and it’s a collection of techniques. Mining is used to eradicate randomness and discover concealed patterns, and it involves searching, retrieving and filtering of relevant data. Data mining can be divided into descriptive mining

and predictive mining, and the techniques of data mining can be divided into three major groups, artificial intelligence technologies, machine learning techniques and statistical techniques. [22]

LLMs' ability to understand and process natural language can allow us to enhance data mining significantly. This enables more accurate data mining, that can extract even more nuanced insights from a vast collection of data. There are a few ways LLMs can enhance data mining, e.g. by enhancing text mining, sentiment analysis and knowledge extraction and creating automated text summarization. As LLMs go beyond identifying just keywords to understanding context and even sentiments within data, they allow for more accurate and efficient text mining. The automated text summarization and knowledge extraction especially are valuable abilities for materials science that LLMs bring to data mining, as they allow for easier and faster way to learn vast amounts of data. This allows for researches to have summaries from huge volumes of text or just specific bits of knowledge that they need at the moment. [23]

3.2 Information Extraction approaches

Majority of scientific knowledge lies across the text, tables and figures in millions of academic papers. Important information for materials science includes e.g. material properties, structure-property relationships and calculations. To be able to extract these, different approaches have been created to replace manual extraction. Main approaches to IE are rule-based, NER-based and LLM-based. Rule-based approaches are based on a set of pre-defined rules to extract information. NER-based approaches use rule-based algorithms to identify relationships between entities and can learn to search for relevant entities within data. LLM-based approaches leverage pre-trained LLMs to understand and extract information. Rule-based ones were the first to be created of these approaches and followed by NER-based. Some of these older approaches are still used alongside the more advanced LLM-based ones. One older method still in use is ChemDataExtractor 2.0. [18], [24]

ChemDataExtractor 2.0 is a rule- and NER-based method for IE, specifically for extracting chemical information from scientific literature [19]. This method extracts unstructured information from PDFs, HTMLs and XMLs and produces an output of machine-readable structured data. ChemDataExtractor 2.0 provides a table processor to extract tabulated experimental properties from the input files. In addition to this, ChemDataExtractor 2.0 uses document-level processing algorithms to resolve data interdependencies and produce unified chemical records. [19]

The first step in extraction with ChemDataExtractor 2.0 is processing files to isolate relevant domains, extract raw text and merge fragmented data from different sources. This leads to consistent document

structure that allows for all documents to be processed in the same way. After this, the documents are processed with a Natural Language Processor, to extract individual chemical records. The extracted information is then processed once again to resolve any interdependencies and the data is combined into an individual structured document that can be deposited to a database. [19]

ChemDataExtractor 2.0 performs well on extracting chemical information from scientific documents, allowing for creation of chemical databases in automated way. To allow for this method to work, it needs a lot of manual setup e.g. writing extraction rules by hand. The method is also domain specific and it doesn't have great flexibility as it would need manual rule updates or retraining when new tasks or formats appear. To address these shortcomings, LLMs are needed. One example of an LLM-based IE tool is ChatExtract.

ChatExtract is a method that can automate very accurate data extraction with minimal initial effort and background [13]. ChatExtract is based on a conversational LLM and engineered prompts, allowing for the LLM to both identify sentences with data and extract the data. In ChatExtract, prompt-engineering is used to automate and increase the accuracy of information extraction achieved. The extraction is done in two main stages: initial classification and a series of prompts. In initial classification a simple relevancy prompt is given to the LLM to provide information whether a sentence contains the data for the property in question. This allows for the elimination of irrelevant sentences and to save time as these sentences don't move to the second stage. [13]

In the second stage of extraction, first the data is split into single- and multi-valued text. This is done as single-valued text is much more likely to be extracted without further prompts whereas multi-valued is more prone for errors and requires further inspection and verification. For single-valued text, simple questions are asked about the text's data, allowing for negative answers to minimize the model's chance of hallucination when enough data isn't presented. If a negative answer is given, the text in hand is discarded and no data is extracted. For multi-valued texts, the model is asked to provide the data in a table in structured form instead. Then each field in the table is inspected and follow-up questions are asked. These follow-up questions are to overcome issues LLMs would have providing false responses without the combination. As for single-valued text negative answers lead to discarding the text, same goes for multi-valued. For multi-valued texts some follow-up questions are uncertainty-inducing to encourage negative answers. This is done so that the model reanalyzes the text instead of reinforcing previous answers. In Table 4 ChemDataExtractor 2.0 and ChatExtract are presented with their different features. [13]

Table 4. IE approaches ChemDataExtractor 2.0 and ChatExtract with their features listed.

ASPECT	CHATEXTRACT	CHEMDATAEXTRACTOR 2.0
BASE	LLM + prompt engineering	rule-based + NER
EXTRACTION	identifies, extracts and verifies materials from research texts via conversational interaction	rule templates, token patterns and supervised NER-models to extract from unstructured data
PERFORMANCE	achieved > 90 % precision on bulk modulus and similar score on other properties	accuracy of 93.4 % for chemical identifiers, 86.8 % for spectroscopic attributes and 91.5 % for chemical properties, lower in diverse sources
USES	rapid, high-accuracy extraction of diverse material properties	extraction of chemical names, properties and spectroscopic data (domain specific)
ADAPTABILITY	high	low/moderate
LIMITS	relies on LLM quality	limited adaptability

LLM-based methods compared to NER-based methods bring more flexibility to IE, as they don't need to be retrained when given new tasks, they can just be given new prompts. This allows for methods that aren't domain specific and can be used for many different domains. As the extraction logic can be defined with prompts instead of coding rules by hand, LLM-based methods reduce development efforts compared to NER-based ones.

3.3 Large Language Models in Information Extraction

LLMs have brought many advancements to IE. They have enabled more accurate and time saving extraction of structured knowledge from unstructured scientific literature and automatic correction of errors. In a study by Lai et al. they found LLM approaches to have significant time savings when compared to conventional ones [25]. Data extraction time was found to have reduced by 80 % when using LLM approaches, compared to conventional manual or rule-based approaches. In the same study they found LLMs to have great extraction performances across different document types and languages. LLM approaches were found to found to have great extraction accuracy, between 95–97 %, often matching or even surpassing conventional approaches' accuracies. [25]

In another study by Dagdelen et al. LLM approaches were found to be efficient in extracting structured data from scientific texts and turn it into structured data [24]. This study also found LLM approaches to be very accurate with extractions. LLM approaches were found to automatically correct errors e.g. correcting the erroneous inclusion of white spaces in chemical formulae. This can be utilized by domain specialists who desire cleaner forms rather than the directly pulled extractions. Using LLMs that have the ability to automatically correct errors lessens the need for correction as a post-processing task. [24]

4 Challenges

Using LLMs has its own risks and challenges on whether the information gained by using LLMs is reliable or is using the models ethical and ecological. One significant challenge is the possibility of hallucinations. Hallucinations are mistakes in the generated text, that are semantically or syntactically plausible but are actually nonsensical or incorrect. This means that the models produce content that is not based on any facts or evidence whereas it's based on the models' own assumptions or biases. [26] Hallucinations in LLMs can be divided into two different categories, intrinsic hallucinations and extrinsic hallucinations. Intrinsic hallucinations introduce logical or factual errors directly opposing the original content. Extrinsic hallucinations include speculative or unconfirmable aspects and cannot be verified against the source. [10]

From an environmental standpoint the use of LLMs is concerning as the development and use of LLMs uses a great amount of electricity and resources. The models are typically built around large-scale neural networks that require large amounts of electricity, and they have immense negative carbon footprints. The carbon emission for training just one generative AI model, e.g. GPT-3, was estimated to be equal to the annual CO_2 emissions of several dozens of households. [26]

The data that is used to train LLMs greatly defines how the models work and what is the quality of the output they have. Concerning this a term called GIGO, "Garbage In, Garbage Out", has been coined, meaning that if the training data is contaminated by e.g. sexist or racist information the output of the model is most likely problematic. This brings ethical and social risks of LLMs that include their prospenity to reproduce harmful biases. Relevancy is also a problem with GIGO, as if the model is trained with outdated data, the output won't be up to date and relevant. In materials science research it's important to have data that is up to date to ensure that any data that isn't up to date, won't sabotage the research. [10]

Wether the LLM that is used is Open Source or Closed Source, brings its own challenges to using it. Open Source LLMs biggest challenge is the risk of misuse. The models can be used to spread misinformation if anyone can just access every aspect of the models. Open Source models also pose a risk for cybersecurity threats, as they can be weaponized to be used in different cybercrimes, e.g. hacking attempts. For Closed Source LLMs, the biggest challenge comes from the lack of visibility and the uncertainty of what kind of biases the models can have. As the models' properties and data are shut from outsiders, users have no way of being absolutely sure about the models' biases or logic for decision making, building uncertainty with using these models. [11]

5 Future directions

As LLMs are still relatively new, a lot of research and development is needed. Few significant directions are development for more sustainability and reliability. A great importance needs to be put on finding more sustainable ways to uphold LLMs as the number of resources needed to train these huge models only increases with development of newer models. Instead of only focusing on developing bigger and better models, we need to think of ways to get their carbon emissions smaller. Sustainable energy producers should collaborate with LLM developers to find solutions that help conserve the planet's resources.

Hallucinations pose a great challenge on the reliability of LLMs. Even though the number of hallucinations can be reduced by rigorous training processes and ensuring the data used is of good quality, it could be developed even more. By doing this we could eventually have LLMs that make few to no hallucinations or mistakes. This would also reduce the need to train models again, if they are noted to make many mistakes. Reliability can also be developed by creating LLMs that have features from both Open Source and Closed Source LLMs.

The capability of LLMs in materials science is determined with both the strength of the models as well as the available tools, hence there being a need to create specific materials science tools for LLMs, e.g. high-throughput computing platforms. [4] As IE with LLMs is largely based on databases, a great importance should be put on expanding the databases and making them accessible for more people. To do this would allow more people to use them and maybe make new innovations.

6 Conclusions

Large Language Models represent a change in materials science in how knowledge is accessed, processed and applied. They enable a more efficient and accurate Information Extraction from data sources, such as publications and web pages, that are unstructured. In a field such as materials science, where most of the used data comes from unstructured sources, LLMs offer a viable and powerful way to do such tasks.

Information Extraction used to be based on rigid rule-based systems but with the development of LLMs, it has evolved significantly. The LLMs abilities to comprehend context, understand nuanced language and structure information with minimal fine-tuning, makes them suited for tasks in IE and especially in materials informatics. ChatExtract is a great example on how LLMs can be utilized for IE. They can be used e.g. for rapid extractions, without compromising the quality and accuracy of the end-results.

Although using LLMs has many great aspects, it doesn't come without challenges. Hallucinations, environmental costs and questions about ethicality and transparency require a great deal of consideration. The quality of the data used in training of LLMs, sets everything that the models do and remains a significant factor on their reliability. As materials science requires a lot of detail and precision, every aspect that can lead to even minor inaccuracies, needs to be assessed carefully.

In the future a key focus should be on developing the reliability and sustainability of LLMs. This entails enhancing transparency and reducing the carbon emissions that come from LLMs. Another key focus should be put on developing databases, to expand them and allow for an open access to high quality data. This development eventually allows for more innovations and interdisciplinary collaborations.

LLMs offer great tools to use already existing information in a way that saves researchers time and resources. Overall, the future of using LLMs in materials science doesn't rely solely on the results we can get, but looking at the whole process and trying to develop it to a more sustainable direction.

References

[1] ‘What Is Artificial Intelligence (AI)? | IBM’. Accessed: Apr. 02, 2025. [Online]. Available: https://www.ibm.com/think/topics/artificial-intelligence

[2] ‘Machine learning (ML): All there is to know’, ISO. Accessed: Apr. 02, 2025. [Online]. Available: https://www.iso.org/artificial-intelligence/machine-learning

[3] M. A. TOR, ‘Top 5 Validation Methods In Machine Learning’, Medium. Accessed: May 14, 2025. [Online]. Available: https://medium.com/@mehmetalitor/top-5-model-validation-methods-in-machine-learning-77eed8d08937

[4] S. Yu, N. Ran, and J. Liu, ‘Large-language models: The game-changers for materials science research’, *Artificial Intelligence Chemistry*, vol. 2, no. 2, p. 100076, Dec. 2024, doi: 10.1016/j.aichem.2024.100076.

[5] OpenAI *et al.*, ‘GPT-4 Technical Report’, Mar. 04, 2024, *arXiv*: arXiv:2303.08774. doi: 10.48550/arXiv.2303.08774.

[6] ‘Natural Language Processing (NLP) Examples | Tableau’. Accessed: May 14, 2025. [Online]. Available: https://www.tableau.com/learn/articles/natural-language-processing-examples

[7] ‘A Brief History of Large Language Models (LLM)’, Parsio Blog. Accessed: Apr. 14, 2025. [Online]. Available: https://parsio.io/blog/a-brief-history-of-llm/

[8] Z. Wang, Z. Chu, T. V. Doan, S. Ni, M. Yang, and W. Zhang, ‘History, development, and principles of large language models: an introductory survey’, *AI Ethics*, Oct. 2024, doi: 10.1007/s43681-024-00583-7.

[9] ‘Next Word Prediction in LLMs’. Accessed: Apr. 29, 2025. [Online]. Available: https://apxml.com/courses/intro-large-language-models/chapter-2-simplified-mechanics-of-llms/predicting-next-word

[10] Y. Annepaka and P. Pakray, ‘Large language models: a survey of their development, capabilities, and applications’, *Knowl Inf Syst*, vol. 67, no. 3, pp. 2967–3022, Mar. 2025, doi: 10.1007/s10115-024-02310-4.

[11] A. Lee, ‘Open Source vs. Closed Source LLMs: The Ultimate Debate’, AI Business Asia. Accessed: May 12, 2025. [Online]. Available: https://www.aibusinessasia.com/en/p/open-source-vs-closed-source-llms-the-ultimate-debate/

[12] ‘Data extraction: Definition, types and more’. Accessed: Apr. 29, 2025. [Online]. Available: https://www.fivetran.com/learn/data-extraction

[13] M. P. Polak and D. Morgan, ‘Extracting accurate materials data from research papers with conversational language models and prompt engineering’, *Nat Commun*, vol. 15, no. 1, p. 1569, Feb. 2024, doi: 10.1038/s41467-024-45914-8.

[14] L. Himanen, A. Geurts, A. S. Foster, and P. Rinke, ‘Data-Driven Materials Science: Status, Challenges, and Perspectives’, *Advanced Science*, vol. 6, no. 21, p. 1900808, 2019, doi: 10.1002/advs.201900808.

[15] Z. Wang *et al.*, ‘Data-Driven Materials Innovation and Applications’, *Advanced Materials*, vol. 34, no. 36, p. 2104113, Sep. 2022, doi: 10.1002/adma.202104113.

[16] 8ai, ‘Databases in Today’s Materials Science - Smart Material’. Accessed: Apr. 03, 2025. [Online]. Available: https://smartmaterial.hi-iberia.es/blog/blog/databases-in-todays-materials-science/

[17] ‘Exploring the World of Materials Databases: A Treasure Trove for Materials Discovery’. Accessed: Apr. 13, 2025. [Online]. Available: https://www.materials.zone/blog/exploring-the-world-of-materials-databases-a-treasure-trove-for-materials-discovery

[18] Z. Huang, P. Peng, F. Lu, and H. Zhang, ‘An LLM-Based Method for Quality Information Extraction From Web Text for Crowed-Sensing Spatiotemporal Data’, *Transactions in GIS*, vol. 29, no. 1, p. e13294, 2025, doi: 10.1111/tgis.13294.

[19] M. C. Swain and J. M. Cole, ‘ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature’, *J. Chem. Inf. Model.*, vol. 56, no. 10, pp. 1894–1904, Oct. 2016, doi: 10.1021/acs.jcim.6b00207.

[20] Y. Hu *et al.*, ‘Improving Large Language Models for Clinical Named Entity Recognition via Prompt Engineering’, Jan. 25, 2024, *arXiv*: arXiv:2303.16416. doi: 10.48550/arXiv.2303.16416.

[21] S. Gupta, A. Mahmood, P. Shetty, A. Adeboye, and R. Ramprasad, ‘Data extraction from polymer literature using large language models’, *Commun Mater*, vol. 5, no. 1, p. 269, Dec. 2024, doi: 10.1038/s43246-024-00708-9.
[22] M. A. Jassim and S. N. Abdulwahid, ‘Data Mining preparation: Process, Techniques and Major Issues in Data Analysis’, *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 1090, no. 1, p. 012053, Mar. 2021, doi: 10.1088/1757-899X/1090/1/012053.
[23] ‘Advanced Data Mining Techniques Using Large Language Models’. Accessed: May 02, 2025. [Online]. Available: https://www.rapidcanvas.ai/blogs/advanced-data-mining-techniques-using-large-language-models
[24] J. Dagdelen *et al.*, ‘Structured information extraction from scientific text with large language models’, *Nat Commun*, vol. 15, no. 1, p. 1418, Feb. 2024, doi: 10.1038/s41467-024-45563-x.
[25] H. Lai *et al.*, ‘Language models for data extraction and risk of bias assessment in complementary medicine’, *npj Digit. Med.*, vol. 8, no. 1, pp. 1–8, Jan. 2025, doi: 10.1038/s41746-025-01457-w.
[26] S. Feuerriegel, J. Hartmann, C. Janiesch, and P. Zschech, ‘Generative AI’, *Bus Inf Syst Eng*, vol. 66, no. 1, pp. 111–126, Feb. 2024, doi: 10.1007/s12599-023-00834-7.