QATAR UNIVERSITY

COLLEGE OF ENGINEERING

# AUTOMATING INFORMATION EXTRACTION FROM PEROVSKITE SOLAR CELLS LITERATURE USING LARGE LANGUAGE MODELS

BY

RADWA ESSAM GAD

A Thesis Submitted to

the College of Engineering

in Partial Fulfillment of the Requirements for the Degree of

Masters of Science in Computing

June 2025

# COMMITTEE PAGE

The members of the Committee approve the Thesis of

Radwa Essam Gad defended on 13/05/2025.

Dr. Tamer Elsayed
Thesis Supervisor

Dr. Yasser Hassan
Thesis Co-Supervisor

Prof. Cagatay Catal
Committee Member

Prof. Saeed Salem
Committee Member

Dr. Tanvir Alam
Committee Member

Dr. Maher Azzouz
Committee Member

Approved:
~~Dean College of Engineering~~

# ABSTRACT

GAD, RADWA, E., Masters : June : 2025, Masters of Science in Computing
Title: Automating Information Extraction from Perovskite Solar Cells Literature Using Large Language Models
Supervisor of Thesis: Dr. Tamer Elsayed.

With the rapid advancement of perovskite solar cells (PSCs) research, efficiently extracting structured data from scientific literature has become essential for accelerating materials discovery and development. PSCs studies often report multiple device configurations within a single paper, making traditional single-device extraction approaches insufficient. In this thesis, we are the first to propose an automated information extraction pipeline that leverages Large Language models (LLMs) to extract structured attributes for all reported devices in PSCs research papers. Our experiments utilize open-source and closed-source LLMs, including GPT-4o-mini, LLaMA 3.1 70B, and Qwen 2.5 72B, ensuring a comprehensive evaluation across various model architectures. Additionally, we introduce the first multi-device evaluation framework using an optimization-based matching algorithm. We also define a wide range of PSC-specific attributes, carefully selected to enhance the practical utility of the extracted dataset for researchers. Our experimental results demonstrate that the proposed pipeline outperforms existing approaches, achieving a champion-device extraction $F_1$ score of 90.06%, $F_1$ score of 78.70% for multi-device extraction, and the best $F_1$ score of 90.98% for the best device in multi-device extraction. These findings highlight the effectiveness of our approach in delivering a scalable, reproducible, and efficient solution for automating structured information extraction from PSCs literature.

# DEDICATION

*To all our people in our beloved Palestine, who have endured oppression and hunger, to the children and youth of Palestine, who carved their dreams onto walls amid the rubble, to a nation unyielding, teaching the world dignity.*
*To my dear mother, father, and sisters, for their unconditional love, endless prayers, motivation, and unwavering support throughout my life, to my friends, Nada and Asmaa, for their encouragement and steadfast support.*

# ACKNOWLEDGMENTS

First and foremost, all praise and gratitude to Almighty Allah for honoring me and bestowing upon me the strength to complete this work. None of this would have been possible without His mercy and countless blessings.

Words cannot fully express my gratitude to my supervisors, Dr. Tamer Elsayed and Dr. Yasser Hassan. Their support, invaluable guidance, and patience have been instrumental in completing this thesis. This work would not have been possible without their continuous encouragement and insightful mentorship.

I am also deeply thankful to Salam Albatarni and Sohaila Eltanbouly for their support and the enriching discussions we shared.

This work reflects the collective support and encouragement I have received, and I am truly grateful to have had such remarkable mentors and colleagues throughout this journey.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# CHAPTER 1: INTRODUCTION

The development of science since the 17th century has relied on the traditional empirical method. As a result, much of the discovery and synthesis of advanced functional materials has been driven by slow, expensive, and often tedious Edisonian trial and error [1], [2]. From the research and development stage to prototyping in laboratories through commercialization, materials processing and characterization can take decades, requiring hazardous chemicals and incurring high costs [2]. The prolonged timeline of conventional material discovery is incompatible with the urgent global push for renewable energy solutions as governments seek accelerated technological advancements to meet climate targets and decarbonization goals. This challenge is particularly evident in the field of solar cells. Since the discovery of the first photovoltaic cell at Bell Laboratories in 1954, despite substantial efforts by the scientific community, solar cells have yet to fully replace non-renewable energy sources, such as petroleum or natural gas.

To bridge this gap, the search for next-generation photovoltaic materials has led to the emergence of Perovskite Solar Cells (PSCs), which have demonstrated unprecedented efficiency gains and potential for low-cost fabrication. Their rapid progress within just over a decade highlights their disruptive potential in the renewable energy landscape. However, a paradigm shift in material discovery and data management is essential to accelerate their path from laboratory breakthroughs to commercial deployment [3], [4]. Leveraging automation, machine learning, and large-scale data integration can significantly enhance our ability to identify stable, high-performance perovskite compositions, ensuring that PSCs make a meaningful contribution to the global energy transition.

Given the remarkable semiconductor properties of metal halide perovskites, the research community has extensively explored their potential for solar cells and other optoelectronic applications, overcoming their instability challenges. However, the rapid pace of discovery—driven by thousands of publications each year—poses a significant challenge for researchers attempting to track and interpret critical advancements. The sheer volume of data makes it nearly impossible for experts to extract meaningful insights or recognize overarching trends without systematic organization. To fully harness the potential of perovskites and accelerate progress in the field, there is an urgent need for comprehensive, structured databases that consolidate and standardize experimental findings, enabling efficient data-driven research.

In 2022, Jacobsson et al. compiled The Perovskite Database, an open-access repository aggregating over 42,000 device performance data points from 15,000 articles published before February 2020 [5]. They constructed their data based on the FAIR data principles, ensuring that the information is Findable, Accessible, Interoperable, and Reusable (FAIR). This approach aims to enhance data transparency, facilitate large-scale analysis, and promote knowledge sharing within the research community. Their work demonstrated how such a database could help filter and visualize device performance trends, as well as conduct large-scale statistical analyses. However, their data extraction method relied entirely on manual efforts, requiring researchers to scan journal articles, extract relevant attributes, and format them according to a predefined schema before database entry. This approach required an estimated 5,000 to 10,000 hours of human labor to process data from 7,400 publications. Additionally, the database expansion relied on voluntary contributions from the research community, resulting in slow updates to the database. For example, between March 2023 and May 2024, the total number of records increased marginally from 43,231 to 43,252, highlighting a lack of sustained input [6]. Furthermore, reliance on human entry introduces formatting inconsistencies and transcription errors, resulting in numerous missing values that complicate data consolidation and limit the database's direct usability. These challenges underscore the need for standardized protocols, unified reporting methods, and automated AI-driven solutions for scientific data extraction [6]–[8].

Given the challenges posed by manual scientific discovery, automation and high-throughput approaches have gained significant traction across multiple domains, including drug discovery, catalysis, Li-ion batteries, and organic solar cells, facilitating intelligent material exploration and enabling a faster transition from fundamental research to commercial applications [3], [4]. In 2013, the Materials Genome Initiative introduced the Materials Project[9], along with other large-scale databases [10]–[15], integrating computational materials datasets with information technology to accelerate discovery. However, these repositories primarily rely on theoretical calculations, while experimentally derived materials data remain scattered across the scientific literature, limiting their applicability. The only large-scale experimental databases currently available mainly focus on inorganic crystal structures rather than comprehensive optoelectronic properties or device performance data [16], [17].

The rapid advancements in Natural Language Processing (NLP) and LLMs have significantly transformed how scientific knowledge is processed and utilized across various disciplines, including materials science [18]. Traditionally, extracting critical information from research papers, patents, and experimental reports has been a manual, time-consuming, and error-prone process, requiring domain experts to carefully sift through vast amounts of unstructured text. NLP techniques, ranging from rule-based methods to deep learning-based language models, have provided automation and efficiency in text mining, classification, and information extraction [19], [20]. More recently, LLMs such as GPT [21] and BERT [22] have demonstrated remarkable capabilities in understanding and generating human-like text, allowing researchers to leverage these models for automated scientific knowledge extraction, summarization, and analysis. These advancements enable the development of large-scale structured databases by extracting key material properties, synthesis conditions, and performance metrics from unstructured literature, facilitating faster discovery and innovation [23].

LLMs are crucial in accelerating data-driven research in materials science by transforming scattered textual data into organized, machine-readable formats. Unlike traditional database curation methods, which often rely on manual annotations or rule-based algorithms, LLMs can interpret complex scientific narratives, extract hidden patterns, and generalize across multiple material science domains with minimal human intervention [24]. This capability is particularly beneficial in PSCs, battery materials, catalysts, and quantum materials, where experimental data is frequently reported in varied and non-standardized formats [25].

Notably, in the field of PSCs, the potential of LLMs to transform traditional research methodologies is immense. Unlike structured databases, scientific papers often detail material properties and experimental results in varied and complex formats, including narrative texts, tables, and figures. LLMs are adept at navigating this diversity, extracting critical data points such as bandgap values, stability conditions, and efficiency metrics from unstructured formats, which are often missed by conventional data extraction tools [26]–[28].

## 1.1. Problem Statement

Despite their potential, LLMs face several challenges when applied to the extraction of materials science information. Scientific knowledge in PSCs is primarily documented in unstructured research papers, making it difficult to systematically extract, analyze, and utilize critical device information. Structured databases are crucial for facilitating large-scale, data-driven research.

To our knowledge, the FAIR database is the only structured database compiled from PSCs' research articles. However, its construction relied entirely on manual data extraction, requiring an estimated 5,000 to 10,000 hours of expert effort to process information from just 8,200 papers. This manual approach poses several challenges. The time-intensive nature of manual data extraction prevents rapid expansion and updates, limiting the database's long-term utility. Manual curation introduces inconsistencies, missing values, and formatting errors, making it challenging to ensure high-quality, standardized data. Expanding the FAIR database or generating similar datasets for different material science subfields remains daunting due to the reliance on human effort. Given these challenges, an automated method for extracting structured data from PSCs literature with minimal human intervention is critical.

LLMs provide a promising avenue for automated information extraction, but several domain-specific challenges remain. Those challenges can be summarized as follows:

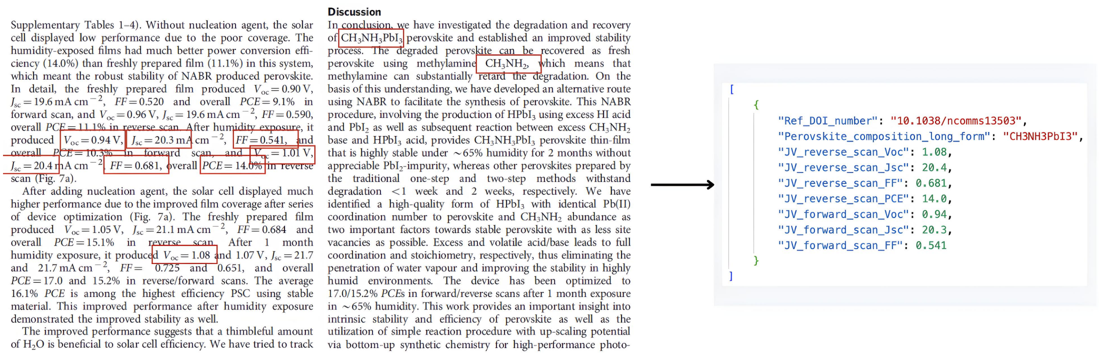
- Attribute Selection: Defining an optimal set of attributes most relevant for PSCs, ensuring that extracted data is valuable and actionable for researchers.

- Multi-Device Extraction: Unlike other material science fields, each research paper on PSCs often describes multiple devices or several device configurations. This introduces extraction and evaluation complexity, as papers report devices with distinct parameters, such as material compositions, deposition methods, and performance metrics. Direct one-to-one comparisons between extracted and ground truth devices become ineffective, requiring an advanced matching strategy to ensure a more reliable evaluation.

- Extraction Strategies: While fine-tuning LLMs can improve extraction accuracy, it is computationally expensive and resource-intensive. Alternatively, prompt engineering presents a lower-cost solution, but its effectiveness in structured PSCs information extraction needs further investigation.

- Evaluation Complexity: Due to the high variability in scientific writing formats, numerical data representations, and implicit reporting styles, evaluating extracted information is non-trivial. In particular, traditional evaluation methods are not well-suited for assessing multi-device extractions, as they struggle to account for the complexity and variability inherent in publications that report multiple device configurations.

To that end, this thesis addresses the problem of automated information extraction and evaluation from PSCs literature, defined as follows: Given a full-text PSCs paper reporting one or more devices, our goal is to extract structured device attributes for all reported devices and evaluate the correctness of the extracted data. Figure 1.1 illustrates an example text paper and the corresponding extracted data.



Supplementary Tables 1–4). Without nucleation agent, the solar cell displayed low performance due to the poor coverage. The humidity-exposed films had much better power conversion efficiency (14.0%) than freshly prepared film (11.1%) in this system, which meant the robust stability of NABR produced perovskite. In detail, the freshly prepared film produced $V_{oc} = 0.90$ V, $J_{sc} = 19.6$ mA cm$^{-2}$, $FF = 0.520$ and overall $PCE = 9.1\%$ in forward scan, and $V_{oc} = 0.96$ V, $J_{sc} = 19.6$ mA cm$^{-2}$, $FF = 0.590$, overall $PCE = 11.1\%$ in reverse scan. After humidity exposure, it produced $V_{oc} = 0.94$ V, $J_{sc} = 20.3$ mA cm$^{-2}$, $FF = 0.541$, and overall $PCE = 10.3\%$ in forward scan, and $V_{oc} = 1.01$ V, $J_{sc} = 20.4$ mA cm$^{-2}$, $FF = 0.681$, overall $PCE = 14.0\%$ in reverse scan (Fig. 7a).

After adding nucleation agent, the solar cell displayed much higher performance due to the improved film coverage after series of device optimization (Fig. 7a). The freshly prepared film produced $V_{oc} = 1.05$ V, $J_{sc} = 21.1$ mA cm$^{-2}$, $FF = 0.684$ and overall $PCE = 15.1\%$ in reverse scan. After 1 month humidity exposure, it produced $V_{oc} = 1.08$ and 1.07 V, $J_{sc} = 21.7$ and 21.7 mA cm$^{-2}$, $FF =$ 0.725 and 0.651, and overall $PCE = 17.0$ and 15.2% in reverse/forward scans. The average 16.1% $PCE$ is among the highest efficiency PSC using stable material. This improved performance after humidity exposure demonstrated the improved stability as well.

The improved performance suggests that a thimbleful amount of $H_2O$ is beneficial to solar cell efficiency. We have tried to track

**Discussion**

In conclusion, we have investigated the degradation and recovery of $CH_3NH_3PbI_3$ perovskite and established an improved stability process. The degraded perovskite can be recovered as fresh perovskite using methylamine $CH_3NH_2$, which means that methylamine can substantially retard the degradation. On the basis of this understanding, we have developed an alternative route using NABR to facilitate the synthesis of perovskite. This NABR procedure, involving the production of $HPbI_3$ using excess HI acid and $PbI_2$ as well as subsequent reaction between excess $CH_3NH_2$ base and $HPbI_3$ acid, provides $CH_3NH_3PbI_3$ perovskite thin-film that is highly stable under ~65% humidity for 2 months without appreciable $PbI_2$-impurity, whereas other perovskites prepared by the traditional one-step and two-step methods withstand degradation <1 week and 2 weeks, respectively. We have identified a high-quality form of $HPbI_3$ with identical Pb(II) coordination number to perovskite and $CH_3NH_2$ abundance as two important factors towards stable perovskite with as less site vacancies as possible. Excess and volatile acid/base leads to full coordination and stoichiometry, respectively, thus eliminating the penetration of water vapour and improving the stability in highly humid environments. The device has been optimized to 17.0/15.2% *PCE*s in forward/reverse scans after 1 month exposure in ~65% humidity. This work provides an important insight into intrinsic stability and efficiency of perovskite as well as the utilization of simple reaction procedure with up-scaling potential via bottom-up synthetic chemistry for high-performance photo-

```
[
    {
        "Ref_DOI_number": "10.1038/ncomms13503",
        "Perovskite_composition_long_form": "CH3NH3PbI3",
        "JV_reverse_scan_Voc": 1.08,
        "JV_reverse_scan_Jsc": 20.4,
        "JV_reverse_scan_FF": 0.681,
        "JV_reverse_scan_PCE": 14.0,
        "JV_forward_scan_Voc": 0.94,
        "JV_forward_scan_Jsc": 20.3,
        "JV_forward_scan_FF": 0.541
    }
]
```

Figure 1.1. Example for the input and output of our approach. The left side shows an excerpt from a PSCs article, where key device attributes such as *Voc*, *Jsc*, *FF*, and *PCE* are highlighted and to be extracted. The right side presents the structured JSON output generated by our approach, capturing the extracted attributes.

To address the problem, we propose an LLM-based pipeline for automated information extraction from the PSCs literature. Our approach has three main components.

The first key component of our approach is **dataset collection and attribute selection**, where we construct a well-defined attribute schema based on expert-driven selection. This ensures that only relevant and high-impact attributes are extracted, making the generated database actionable and meaningful for researchers.

The second component is the **information extraction**, which encompasses both single and multi-device extraction, enabling the identification and extraction of multiple device configurations per research paper. This feature addresses a significant limitation in prior work, where most extraction efforts focused only on single-device data extraction. Since PSCs studies frequently report multiple devices per publication, our framework is designed to capture this complexity, ensuring a more complete and accurate representation of the dataset.

Finally, to ensure the reliability of the extracted data, we developed a robust **evaluation framework** that accounts for single- and multi-device extraction evaluations.

## 1.2. Objectives

This work aims to:

- Automate the extraction of structured PSC information from full-text research papers using LLMs.
- Develop a scalable and reproducible dual extraction framework that supports processing single and multiple devices per publication.
- Develop an advanced evaluation methodology that evaluates the effectiveness of extracting single and multiple devices.

- Construct a high-quality, standardized dataset of extracted PSC attributes from full research papers, improving accessibility and usability for researchers.

## 1.3. Contributions

Our work presents several contributions:

- Our proposed approach outperforms the state-of-the-art rule-based and LLM-based approaches.
- We are the *first* to introduce a structured approach to extract multiple PSCs devices per publication, making it possible to automate a database construction in a scalable and reproducible manner.
- We are the *first* to propose an evaluation framework that can evaluate both single and multiple devices.
- We selected a comprehensive set of attributes, focusing on key performance metrics (e.g., Current Voltage (JV) parameters, Power Conversion Efficiency), ensuring high relevance for PSCs researchers.
- We systematically evaluate zero-shot, few-shot, and CoT prompting, demonstrating that prompting techniques can achieve competitive performance relative to fine-tuning.
- While focused on PSCs, our modular extraction framework is highly adaptable. It can be applied to various subfields in materials science, enabling the automated extraction of knowledge across diverse materials science domains.

## 1.4. Research Questions

To systematically explore the effectiveness of LLMs in PSCs extraction and evaluation, this study aims to answer the following research questions:

**RQ1:** How do different prompt engineering techniques (few-shot, zero-shot, CoT) impact extraction performance?

**RQ2:** How does extracting all devices affect overall extraction performance compared to focusing on the champion device only (the device with the highest Power Conversion Efficiency (PCE) of their JV reverse scan)?

**RQ3:** Is fine-tuning more effective than extensive prompt engineering for material science information extraction?

**RQ4:** How does our approach compare to state-of-the-art (SOTA) methods?

The remainder of this thesis is structured as follows. Chapter 2 discusses related work. Chapter 3 describes our proposed methodology. Chapter 4 illustrates the experimental evaluation and discusses the results. Chapter 5 provides concluding remarks and presents potential future directions for this work.

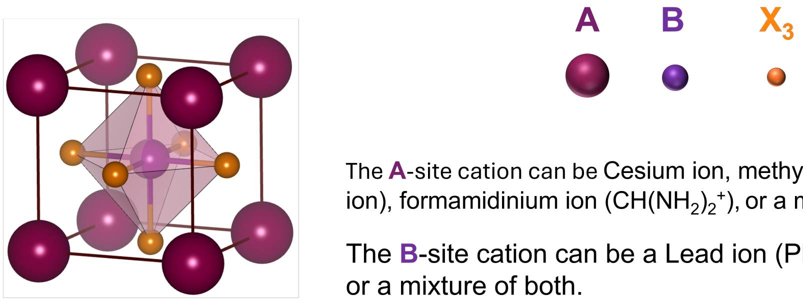
# CHAPTER 2: RELATED WORK

## 2.1. State-of-the-art Material for Solar Cells

Since the advent of silicon (Si) solar cells in the 1950s, photovoltaics have undergone significant evolution through multiple generations of materials and technologies. First-generation solar cells, dominated by crystalline silicon, laid the foundation for commercial photovoltaics but faced limitations in terms of cost and efficiency. Second-generation thin-film technologies, including cadmium telluride (CdTe) and copper indium gallium selenide (CIGS), sought to reduce material usage and manufacturing costs while maintaining competitive performance. However, the emergence of third-generation photovoltaics—particularly PSCs revolutionized the field by achieving rapid efficiency gains, low-cost fabrication, and tunable optoelectronic properties. Unlike traditional semiconductor materials, perovskites possess defect-tolerant electronic structures and high absorption coefficients, enabling efficiencies exceeding 26% in just over a decade of research. Moreover, their potential to surpass the Shockley-Queisser limit through tandem integration with silicon or all-perovskite architectures makes them a state-of-the-art material poised to redefine the future of solar energy.

**A state-of-the-art material:** Perovskites, an emerging class of organic-inorganic hybrid semiconductors made from metal halides, are outstanding candidates to revolutionize both power consumption and production [29], [30].

Halide perovskites has the general formula $ABX_3$, feature a cubic structure where **A** is a larger monovalent cation (e.g., cesium ion ($Cs^+$, methylammonium ($CH_3NH_3^+$) ion), formamidinium ion ($CH(NH_2)_2^+$), or a mixture of these), **B** is a smaller metal cation (e.g., $Pb^{2+}$ or $Sn^{2+}$). **X** is a halide anion ($Cl^-$, $Br^-$, or $I^-$). Their unique crystal structure, consisting of corner-sharing $BX_6$ octahedra with A-site cations filling the interstitial spaces, yields exceptional optoelectronic properties.



The A-site cation can be Cesium ion, methylammonium ($CH_3NH_3^+$) ion), formamidinium ion ($CH(NH_2)_2^+$), or a mixture of these.

The B-site cation can be a Lead ion ($Pb^{2+}$), a Tin ion ($Sn^{2+}$), or a mixture of both.

The X-site can be occupied by halide ions such as chloride ($Cl^-$), bromide ($Br^-$), iodide ($I^-$), or their mixture."

Figure 2.1. Perovskite lattice (one unit cell structure): the solid red line denotes the cubic structure.

The energy efficiency [29] of this class of semiconductors is due to their extraordinary properties such as having a panchromatic absorption profile, exhibiting intense and narrow-band luminescence (strong absorption coefficient of $\sim\ 105\ cm^{-1}$), and possessing excellent ambipolar charge carrier mobilities as well as relatively long carrier diffusion length ($> 1\ \mu m$), bringing them to the forefront of emerging optoelectronic materials (e.g. PVs [30]–[35], LEDs [36]–[40] and electroluminescent devices [41]). Despite these promising features, the commercialization of metal halide perovskites faces significant challenges that affect their performance metrics and eventually lead to degradation trails, putting the device's long-term stability at risk [42], [43].

## 2.2. NLP for Material Science

NLP has played an increasingly significant role in materials science and materials discovery by enabling the automated extraction of information, text mining, classification, and analysis of scientific literature [18]. Early NLP applications in materials science primarily relied on rule-based approaches and supervised learning models [44]–[46]. However, the advent of transformer-based architectures such as BERT (Bidirectional Encoder Representations from Transformers) [22] and GPT (Generative Pretrained Transformer) [21] has significantly enhanced the ability of models to understand domain-specific terminology, extract relationships between material properties, and generate meaningful predictions. This section reviews key advancements in NLP for materials science, focusing on BERT-based models and LLMs such as GPT, as well as domain-adapted models.

### *2.2.1. BERT-Based Models for Materials Science*

BERT and its domain-specific variants have been widely adopted in materials science applications due to their robust contextual understanding and adaptability for text mining. In materials science literature, several studies have pre-trained and fine-tuned BERT-based models to improve named entity recognition and information extraction tasks.

One notable adaptation is MatBERT, which was trained on materials science corpora. [47] demonstrated that MatBERT outperforms general-purpose BERT and even SciBERT [48], highlighting the importance of domain-specific pre-training for improving named entity recognition (NER) tasks in materials science literature. Their study also found that even a simpler BiLSTM model, trained with materials science embeddings, outperformed general-purpose BERT models, emphasizing the necessity of fine-tuned pre-training.

Another specialized model, MatSciBERT, was developed as a domain-specific model for material science to further refine entity recognition and relation extraction [49].

Furthermore, BatteryBERT was introduced as a domain-specific NLP model tailored for research on battery materials [50]. The study demonstrated that models like BatteryBERT significantly outperform general BERT models in extracting relevant material properties, synthesis methods, and applications from research papers.

In addition, OpticalBERT was designed for text and table-based language modeling within the optical materials domain [51]. This work highlights how BERT-based models can be tailored to support specific material science subfields, thereby enhancing structured information extraction.

MatSciNLP was developed as a benchmark to evaluate NLP models on various materials science tasks, such as named entity recognition, relation classification, and synthesis action retrieval [52]. This benchmark highlighted the benefits of using pretrained domain-specific models, particularly MatBERT, which outperformed general-purpose models in materials science text-processing tasks.

A crucial finding across these models is that fine-tuning on domain-specific corpora significantly improves performance in materials science tasks. However, despite these advancements, BERT-based models still face challenges in handling long-form scientific documents, complex multi-device descriptions, and extracting implicit knowledge.

### *2.2.2. Large Language Models for Materials Science*

The advent of large generative models, such as GPT-based models, LLaMA, and domain-specific LLMs, has provided new opportunities for materials science, surpassing traditional BERT-based approaches in handling complex language structures and reasoning-based tasks.

Choi et al. [53] proposed a GPT-enabled materials language processing (MLP) pipeline for chemistry and materials science applications, demonstrating that prompt-based approaches can achieve comparable accuracy to fine-tuned BERT models in document classification, named entity recognition, and extractive question answering. Their work highlights the efficiency of strategic prompt engineering in replacing complex architectures traditionally required for materials information extraction.

Recent studies have evaluated GPT-4's capabilities in chemical and materials science research, revealing its potential for foundational chemistry knowledge, cheminformatics, data analysis, and hypothesis generation [54].

ChemLLM [55], a chemical large language model, demonstrated strong performance in chemical property prediction and material synthesis reasoning. Similarly, Choi et al. [53] proposed a GPT-enabled MLP pipeline for chemistry and materials science applications, demonstrating that prompt-based approaches can achieve comparable accuracy to fine-tuned BERT models in document classification, named entity recognition, and extractive question answering. Their work highlights the efficiency of strategic prompt engineering in replacing complex architectures traditionally required for materials information extraction. The work found that while these models excel at general scientific understanding, they often struggle with the specificity required for materials science tasks, such as parsing precise compositions and extracting synthesis procedures.

The DARWIN SERIES introduced domain-specific LLMs for natural science, incorporating progressive instruction fine-tuning to improve model adaptation to complex material descriptions [56]. Similarly, MatChat [57], a large language model and service platform for materials science, was designed to provide real-time assistance for materials research queries, showcasing how conversational AI can enhance information retrieval in scientific literature.

Beyond retrieval and extraction tasks, LLMs have also been explored for generating scientific hypotheses in materials science. Studies have assessed whether models like ChatGPT can generate novel hypotheses by reasoning through existing literature [58], [59]. For instance, [58] demonstrated that ChatGPT could assist in hypothesis generation for improving perovskite solar cells. Their study used ChatGPT to identify polyallylamine (PAA) as a potential surface modifier, a molecule that was subsequently experimentally validated to enhance device efficiency.

Despite their advantages, LLMs introduce challenges related to hallucination, domain adaptation, and evaluation, requiring rigorous validation methods to ensure reliability. The following sections explore how these models are applied to information extraction and evaluation within materials science.

## 2.3. Information Extraction in Material Science

Information extraction in materials science is crucial for building structured databases from unstructured text, enabling efficient access to scientific knowledge. Traditionally, IE has relied on rule-based NLP techniques, named entity recognition (NER), and algorithmic approaches to extract relevant information. However, these methods often require extensive domain expertise, manual rule-setting, and predefined heuristics, limiting their scalability and adaptability to new research.

### *2.3.1. Information Extraction in PSCs*

A rule-based NLP approach was employed by Valencia et al. [6], who developed an NLP-driven system for extracting the fabrication details of PSCs from journal articles. Their method utilized NLP tools, such as ChemDataExtractor, alongside algorithmic techniques to systematically extract device data, extrinsic cell definitions, and fabrication procedures. By processing 3,164 journal articles, their system achieved an average extraction accuracy of 89.9%, demonstrating the effectiveness of structured algorithmic methods in automating database generation. However, while this approach proved reliable, it still required carefully designed extraction rules and was limited in its ability to generalize across different material domains.

Similarly, Zhang et al. [60] developed an NLP-driven framework for exploring literature and discovering PSC materials. Their model utilized the word2vec technique to analyze 29,060 scientific publications, successfully learning key concepts such as light-absorbing, electron-transporting, and hole-transporting materials in PSCs. The NLP model identified a novel hole-transporting material ($Fe_3O_4$), which was subsequently validated through density functional theory (DFT) calculations and device experiments.

With recent advances in LLMs, IE has transformed, reducing its dependency on manual rule-setting and enabling more flexible data extraction with minimal human intervention. Xie et al. [7] introduced a structured information inference (SII) framework that leveraged fine-tuned LLaMA models to convert unstructured data on Perovskite solar cells materials into structured formats. Their approach employed a pipeline to extract and standardize key material attributes directly from research papers, achieving an $F_1$ score of 87.14%. Unlike rule-based NLP methods, the model demonstrated the ability to generalize across diverse material descriptions and could further support the predictive modeling of material properties.

### *2.3.2. Information Extraction Across Material Science Domains*

Beyond perovskite solar cells, efforts have been made to extract data from a broader range of material science subfields.

Dagdelen et al. [23] presented an approach that fine-tunes pre-trained large language models (such as GPT-3 and LLaMA-2) to perform NER and relation extraction jointly. Their method extracts structured knowledge from scientific literature, generating output in plain text or structured formats, such as JSON. This approach was tested on materials chemistry-related tasks, demonstrating significant improvements in accuracy and usability compared to traditional IE methods. The study highlights the potential of LLMs to create large, structured scientific knowledge bases from unstructured research papers, which is directly relevant to our work on extracting structured information from the perovskite field.

Another study by Gupta et al. [61] investigated the capacity of GPT-based models to extract reaction conditions from polymer literature, demonstrating that fine-tuned LLMs significantly outperform general-purpose models. Their approach achieved a 78.6% accuracy in extracting polymer reaction data, highlighting the benefits of domain-specific fine-tuning.

Polak Morgan (2024) [62] introduced ChatExtract, a method that leverages conversational LLMs to extract material data with high precision, achieving 90.8% precision and 87.7% recall for bulk modulus extraction and critical cooling rates of metallic glasses.

Similarly, Leong et al. [63] proposed a multimodal LLM-based reaction mining pipeline (MERMES) to extract chemical reaction data from text, figures, and tables, demonstrating 96% accuracy in reaction parsing.

In computational material science, MatScIE [64] was developed as an automated tool for generating structured databases from computational materials literature. The system extracted methods, parameters, and results from papers, streamlining the knowledge extraction process for computational materials studies. Their approach demonstrated that automated NLP pipelines could facilitate large-scale data aggregation and reduce the reliance on manual curation.

A comprehensive study by Wang et al. [26] examined the performance of generative LLMs on domain-specific information extraction, specifically in extracting bandgap data from materials science literature. The study compared GPT-4 against a rule-based extraction method (ChemDataExtractor [19]) and found that GPT-4 significantly outperformed the rule-based approach, achieving an 87.95% correctness rate. In contrast, the rule-based method reached only 51.08% correctness. Notably, their evaluation was conducted manually by human experts, who assessed the correctness of each extracted bandgap value. This further motivated the development of robust and advanced evaluation strategies that can provide consistent and scalable assessments for similar tasks.

## 2.4. Prompt Engineering in Material Science

Prompt engineering has emerged as a crucial technique for optimizing LLM performance in scientific data extraction. Instead of retraining or fine-tuning models, researchers can craft structured prompts to guide LLMs toward more accurate and context-aware outputs.

Chen et al. [65] examined zero-shot and few-shot prompting techniques for extracting chemical-disease relations. Their work found that few-shot prompting outperformed rule-based methods by 16.7% in $F_1$ scores, demonstrating the effectiveness of example-based prompting strategies. Xia et al. [66] refined prompt engineering for systematic literature reviews, showing that LLM-generated screening results reduced human workload by over 80%.

Similarly, Polak and Morgan [62] developed ChatExtract, a prompt-based workflow for extracting materials data. Their study showed that by applying carefully engineered follow-up prompts, ChatExtract achieved 90.8% precision and 87.7% recall in extracting bulk modulus data while also maintaining high accuracy (91.6% precision, 83.6% recall) in extracting critical cooling rates for metallic glasses. The authors demonstrated that introducing uncertainty-inducing redundant questioning in prompts significantly reduced errors and hallucinations in extracted data.

LLMs have also been employed to classify materials through prompt engineering. Liu et al. [67] developed a deep learning workflow that combines LLM-generated textual features with a BERT-based classifier to improve classification accuracy. Their approach was tested on metallic glass classification, achieving up to a 463% increase in classification accuracy compared to traditional machine learning methods. The workflow demonstrated the potential of prompt engineering to distill scientific knowledge from LLMs and transform it into structured labels for material databases.

A key challenge in leveraging LLMs for materials science lies in integrating domain-specific knowledge into their reasoning processes. Liu et al. [68] introduced a domain-knowledge-embedded prompt engineering framework to enhance LLM performance in chemistry and materials science. Their study demonstrated that incorporating structured domain knowledge into prompts significantly improved capability, accuracy, and $F_1$ scores across multiple tasks, including materials classification and chemical property prediction. The approach also led to a notable reduction in hallucinations, ensuring more reliable outputs.

Another challenge in materials science is the interpretation of phase diagrams, which are critical for understanding material stability and processing conditions. A recent study explored prompt tuning strategies for LLMs to improve their ability to comprehend and analyze phase diagrams [69]. The study found that carefully designed few-shot prompting strategies significantly improved the precision of phase classification tasks.

# CHAPTER 3: METHODOLOGY

This chapter presents the approach followed in our work, detailing the methods used for extracting materials science information. It describes the dataset, the prompt engineering techniques, the fine-tuning process, and the strategies employed to evaluate the performance of the extraction.

## 3.1. Problem Definition

This thesis tackles the challenge of automating information extraction and evaluation from PSC literature. Specifically, given a full-text PSCs research paper that describes one or more devices, our objective is to extract a well-defined set of structured device attributes (e.g., JV reverse scan PCE and Bandgap) for each device reported in the paper and assess the efficiency of the extracted data.

## 3.2. Approach Overview

To address the problem, we propose an approach that systematically extracts structured data from PSCs literature. Our methodology consists of three main phases, as illustrated in Figure 3.1.

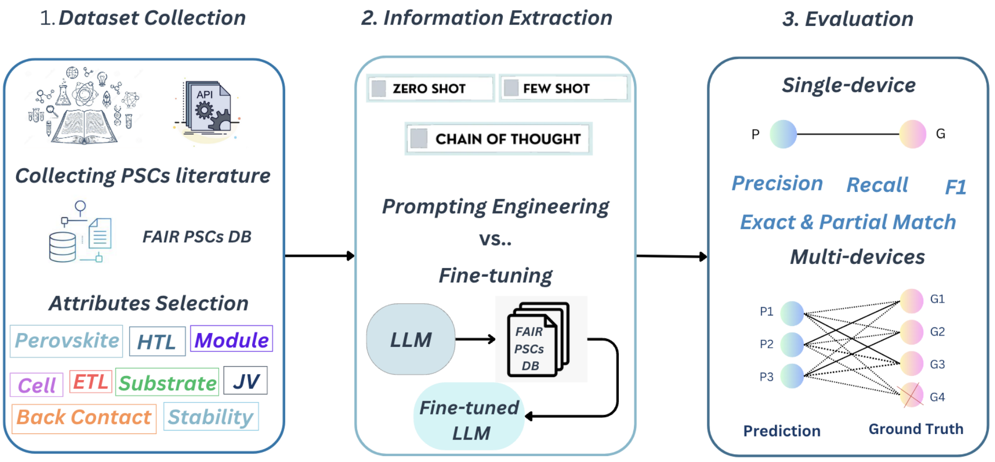


Figure 3.1. Illustration of the proposed methodology

The first phase is dataset collection and input processing, where full research papers related to perovskite solar cells are collected from multiple publication venues. These collected documents are preprocessed and converted to plain text to ensure they are suitable for further processing.

The second phase focuses on information extraction using LLMs. We explore two distinct strategies for extracting scientific information. The first strategy is prompt engineering-based extraction, which does not require prior structured or annotated data. Instead, it relies on various prompt engineering techniques to guide the LLMs in extracting relevant information. The second strategy is fine-tuning-based extraction, where LLMs are fine-tuned on domain-specific, annotated data. This method enables models to learn domain-specific knowledge and enhance extraction performance. The extracted information is then stored in a structured JSON format, following a predefined schema.

The final phase involves our proposed evaluation methodology. Our evaluation methodology assesses the performance of various extraction approaches, encompassing both single-device and multi-device scenarios.

## 3.3. Dataset Construction

Instead of using summaries or abstracts, our approach processes entire research papers as input to assess the capability of LLMs in extracting structured data from long-form scientific content. With advancements in LLM architectures, many models now support extended contexts of over 128K tokens (e.g. Llama-3.1[1] and GPT-4o-mini[2]), making them well-suited for handling extensive scientific documents. This ability is particularly beneficial for our problem and similar challenges, where extracting structured data from lengthy research papers is essential. This enables us to assess how effectively LLMs can handle complete scientific documents without prior segmentation or summarization.

The construction of the dataset for this work involves two primary steps: collecting related research papers and selecting device-specific attributes.

### *3.3.1. Data Collection*

To build a comprehensive dataset for information extraction in PSCs, we collected research papers from multiple sources with API access: Elsevier[3], Springer[4], CORE[5], and arXiv[6]. Furthermore, we used SciCrawler [56], a web scraping tool, to collect data from other journals, including the Royal Society of Chemistry (RSC)[7].

We utilized domain-specific search terms selected carefully by a materials science expert to ensure comprehensive coverage of relevant papers. The search terms included are as follows:

> ”Lead halide PSCs,” ”Lead halide optoelectronics,” ”halide segregation,” ”Perovskite Solar Cells Power Conversion Efficiencies,” ”Perovskite Thin Films,” ”perovskite solar cells and Electron Transport Layer”, ”perovskite solar cells and hole Transport Layer,” ”Tandem Solar Cells”, ”Organic-inorganic perovskites,” ”lead-free perovskites,” ”perovskite photovoltaics,” ”Double Perovskites,” ”Hysteresis,” ”Halide Segregation in Mixed-Halide Perovskites,” ”perovskite solar cells stability.”

Although we collected around 120,000 papers, annotating such a vast dataset is a labor-intensive and time-consuming process that requires significant expertise and resources. To address these challenges, we sourced an existing annotated dataset that closely aligns with our research focus.

Consequently, we leveraged the FAIR Database [5] as our ground truth reference, which includes structured data on perovskite solar cells from over 42,000 devices manually collected and extracted by 91 domain experts from approximately 7,400 papers.

[1] https://github.com/meta-llama/llama-models/tree/main/models/llama3_1

[2] https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

[3] https://dev.elsevier.com/

[4] https://dev.springernature.com/

[5] https://core.ac.uk/services/api

[6] https://info.arxiv.org/help/api

[7] https://developer.rsc.org/

Due to the unavailability of all FAIR papers in our collected corpus, we applied two filtering criteria:

- The paper must be in the FAIR database and our collected dataset.
- The paper should report between 1 and 4 devices to ensure that the most relevant device information is contained within the main paper text rather than in supplementary materials.

From the 2,884 papers available in FAIR, 1,325 matched both criteria, forming the final dataset used in our experiments.

### *3.3.2. Attribute Selection*

In this study, we selected 77 key parameters from the original 410 chosen in the FAIR dataset to ensure a focused, high-impact dataset that captures the most critical factors influencing perovskite solar cell efficiency and stability. Our selection prioritizes parameters that (i) directly affect device performance, such as deposition procedures, material compositions, and photovoltaic metrics; (ii) are commonly reported in the literature, reducing inconsistencies and algorithmic errors when analyzing large datasets; and (iii) enable comprehensive trend analysis, facilitating the identification of research gaps and guiding the development of more efficient PSCs and modules. Limiting the dataset to these essential attributes allows us to balance completeness with practicality, ensuring reproducibility while streamlining data management and interpretation.

To better illustrate the categorization of data in our selected dataset, Figure 3.2 provides an overview of the distribution of attributes across different perovskite-related parameters. This visualization highlights the structured classification of key features, ensuring a well-balanced and representative dataset.

The following concisely justifies why our shortlist of 77 attributes is more focused and practical than the 410 attributes in the FAIR dataset.

- **Core Experimental Focus:** Our selection concentrates on essential parameters—deposition procedures, layer compositions, and key performance metrics (JV, stability, EQE)—most relevant for assessing perovskite device performance. In contrast, the full FAIR set includes extensive ancillary details (i.e., reagent suppliers and solvent purity) that may not impact reproducibility or comparative studies.
- **Reduced Redundancy:** The FAIR dataset lists multiple overlapping details (i.e., separate entries for deposition atmospheres, solvent mixing ratios, and environmental conditions across various layers). In contrast, our shortlist consolidates information to avoid duplication. This streamlining minimizes the risk of inconsistencies during data entry and analysis [70].
- **Enhanced Usability and Clarity:** our attribute set facilitates more precise data analysis and interpretation by focusing on key process and performance parameters. This balance between compositional details and performance outcomes ensures that users can more directly correlate fabrication parameters with device metrics—a critical factor for reproducibility and meta-analysis.

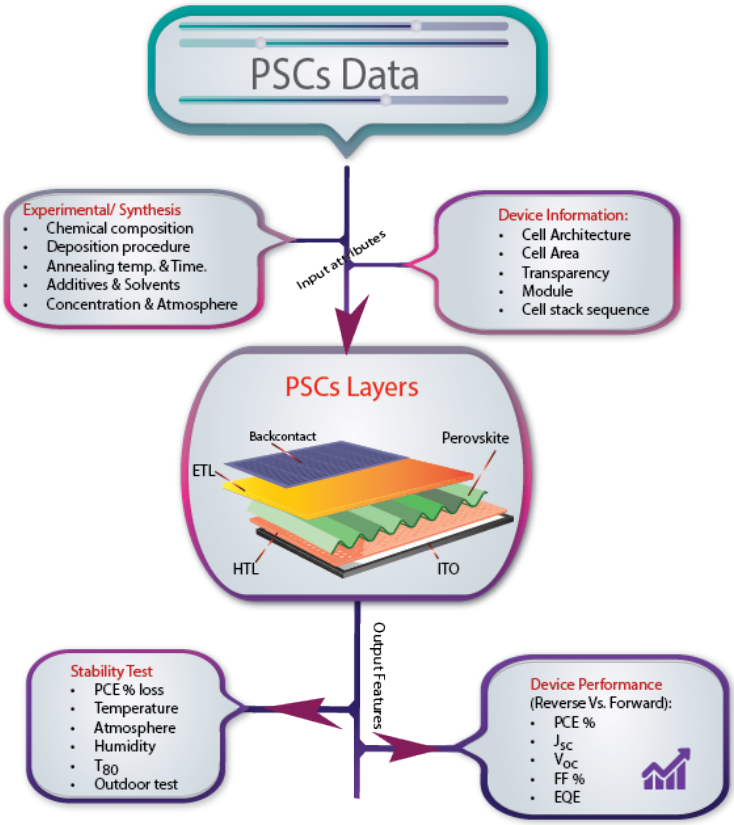


Figure 3.2. Categorization of our 77 selected attributes across different Perovskite-related parameters.

- **Practicality for Routine Reporting:** While the FAIR dataset is designed to be exhaustive for archival purposes, our shortlist is tailored for routine reporting and comparative studies, ensuring that researchers are not burdened with extraneous details. This efficiency is particularly beneficial when integrating data from multiple sources or conducting high-throughput studies.

## 3.4. Information Extraction

In our work, we applied an information extraction approach that utilizes LLMs with prompt engineering and fine-tuning techniques to extract key attributes related to PSCs.

### *3.4.1. Prompt Engineering*

Prompt engineering plays a crucial role in optimizing LLM outputs by strategically designing inputs that guide the model toward producing structured, contextually relevant responses [71]. Well-crafted prompts significantly enhance their ability to extract information accurately. To systematically evaluate the impact of prompt engineering on our task, we experimented with three distinct strategies: Zero-Shot, Few-Shot, and Chain-of-Thought (CoT) prompting. Each of these methods offers distinct advantages and helps mitigate various challenges associated with extracting information from scientific literature. Our prompt engineering approach follows the guidelines provided by OpenAI[8] to ensure clarity, precision, and structured reasoning in extracting complex scientific attributes.

A critical consideration in prompt design is clear and unambiguous instruction. LLMs tend to perform better when they are explicitly directed toward the desired format, scope, and level of detail. Furthermore, we ensured that our prompts adhered to best practices such as delimiting key sections, specifying reasoning steps, and providing structured output formats.

To ensure consistency across different runs, we designed all prompts to output structured data in a predefined JSON schema, maintaining uniformity in extracted attributes. This approach helps reduce variance in responses and simplifies post-processing tasks. Our experiments were conducted using three LLMs: GPT-4o-mini[9], LLaMA-3.1-70B[10], and Qwen-2.5-72B[11]. This ensured that insights drawn from prompt engineering generalize across multiple architectures.

To evaluate the impact of different prompting techniques, we conducted these experiments using the development (dev) set. This dataset was selected to allow iterative testing and refinement of prompt structures without influencing the final evaluation results. Using the dev set, we ensured that the observed performance improvements were not biased by exposure to the test set, maintaining a clear separation between prompt tuning and final model evaluation.

### *3.4.1.1. Zero-Shot Prompting*

Zero-shot prompting is a minimal-intervention approach where the model is provided with task instructions but no specific examples. This method tests the pre-trained capabilities of the LLM in understanding and extracting domain-specific attributes without additional guidance.

**You are an expert in extracting structured data from material science papers, specifically the PSCs.**
*You will receive a full paper in text format and a JSON schema.*
Your task is to extract information on **solar cell stack** and **method information** for **all device configurations** described in the paper.
Strictly follow the schema and output a valid JSON object.
If the paper describes multiple device configurations, output them as an array of JSON objects, where each object follows the provided schema.
**The output should be in the following format:**

```
[ {{'Ref_DOI_number': "doi1", "device": 1, ...}},
{{'Ref_DOI_number': "doi1", "device": 2, ...}} ],
{{'Ref_DOI_number': "doi2", "device": 1, ...}}
```

**Follow these rules:**
1. **Carefully read** the paper and identify all relevant details that match the schema fields.
2. **Ensure each device configuration** is represented as a separate JSON object with a unique **"Device"** number.
3. If a field cannot be determined from the text, set its value to **"Unknown"** or an empty string ("") as specified in the schema.
4. Ensure the output JSON object includes all fields from the schema, even if some are empty or unknown.
5. Do not add or remove any fields from the schema.
6. Format the output as a valid JSON object with proper indentation.

**Provide your output strictly following the same structure for any new input, separating the devices.**

Figure 3.3. Zero-Shot Extraction Prompt

A major advantage of zero-shot prompting is its efficiency, since no labeled examples are required, it can be applied to unseen tasks without extensive fine-tuning.

[8] https://platform.openai.com/docs/guides/prompt-engineering

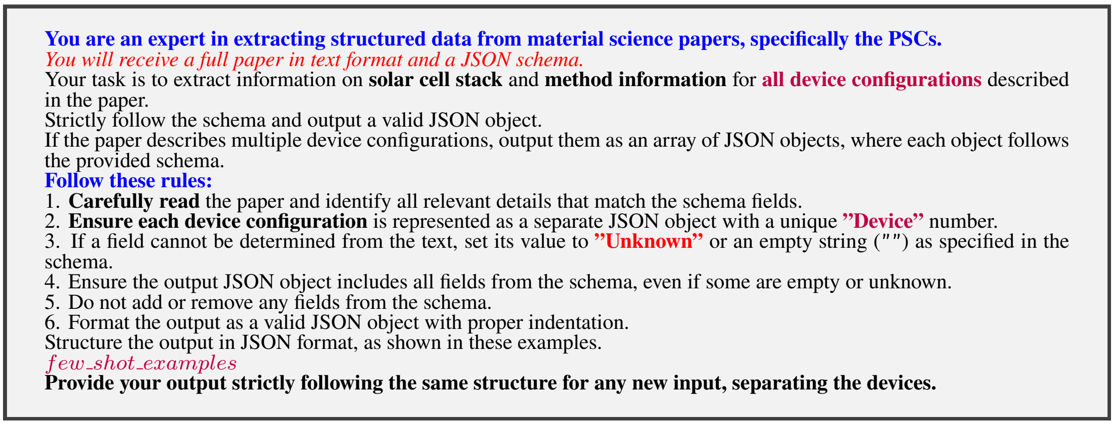
[9] https://openai.com/

[10] https://www.llama.com/

[11] https://www.alibabacloud.com/product/qwen

However, the primary challenge with this approach is the high variability in responses, as the model may misinterpret vague instructions or generate outputs that are inconsistent in format. In many cases, zero-shot prompting is prone to hallucinations, where the model fills in missing details based on prior knowledge rather than strictly adhering to the provided text.

We attempted to mitigate this by refining our prompt to explicitly instruct the model to extract only the information present in the document, leaving unknown fields as ”Unknown” rather than speculating. This adjustment significantly improved the reliability of extracted attributes. Figure 3.3 presents the prompt used in our zero-shot experiments.



**You are an expert in extracting structured data from material science papers, specifically the PSCs.**
*You will receive a full paper in text format and a JSON schema.*
Your task is to extract information on **solar cell stack** and **method information** for **all device configurations** described in the paper.
Strictly follow the schema and output a valid JSON object.
If the paper describes multiple device configurations, output them as an array of JSON objects, where each object follows the provided schema.
**Follow these rules:**
1. **Carefully read** the paper and identify all relevant details that match the schema fields.
2. **Ensure each device configuration** is represented as a separate JSON object with a unique **”Device”** number.
3. If a field cannot be determined from the text, set its value to **”Unknown”** or an empty string ("") as specified in the schema.
4. Ensure the output JSON object includes all fields from the schema, even if some are empty or unknown.
5. Do not add or remove any fields from the schema.
6. Format the output as a valid JSON object with proper indentation.
Structure the output in JSON format, as shown in these examples.
*few_shot_examples*
**Provide your output strictly following the same structure for any new input, separating the devices.**

Figure 3.4. Few-Shot Extraction Prompt

### *3.4.1.2. Few-Shot Prompting*

Few-shot prompting extends the zero-shot approach by incorporating demonstrations—that is, a small set of manually curated examples—within the input prompt. By seeing examples of correctly formatted extractions, the model gains a stronger understanding of expected outputs, reducing ambiguity and improving structured consistency.

The effectiveness of few-shot prompting stems from its ability to provide an implicit training signal within the prompt itself. Instead of relying only on prior knowledge, the model can align its responses with patterns present in the examples. We observed that few-shot prompting significantly improved precision and recall, particularly for attributes requiring multi-word extractions (e.g., chemical compositions, deposition procedures).

One challenge with few-shot prompting is that the quality of selected examples heavily influences performance. Poorly chosen or ambiguous demonstrations may introduce biases or reinforce incorrect predictions. To ensure robustness, we carefully chose examples that capture most of the cases, specifically, samples that cover single and multiple devices. In this experiment, we used the same base prompt from the zero-shot setup but supplemented it with a few representative samples to guide the model’s responses. Figure 3.4 presents the prompt used in our few-shot experiment.

**You are an expert in extracting structured data from material science papers, specifically, the Perovskite solar cells.**
*You will be provided with a full paper in text format and a JSON schema.*
Your task is to extract information on **solar cell stack** and **method information** for **all device configurations** described in the paper.
Strictly follow the schema and output a valid JSON object.
If the paper describes more than one device configuration, output them as an array of JSON objects, where each object follows the provided schema.
**The output should be in the following format:**

```
[ {{'Ref_DOI_number': "doi1", "device": 1, ...}},
{{'Ref_DOI_number': "doi1", "device": 2, ...}} ],
{{'Ref_DOI_number': "doi2", "device": 1, ...}}
```

**Follow these rules:**
1. **Identify the Substrate Stack Sequence**: Extract information about the substrate layers (e.g., ”SLG — ITO”) and their cleaning procedures.
2. **Extract ETL Details**: Identify details about the electron transport layer (ETL), including stack sequence, additives, deposition methods, solvents, and thermal annealing conditions.
3. **Extract Perovskite Details**: Find details on perovskite composition, additives, thickness, deposition methods, solvents, and any surface treatments.
4. **Extract HTL Details**: Locate the hole transport layer (HTL) stack sequence, additives, deposition methods, and thermal annealing conditions.
5. **Extract Back Contact Details**: Identify the back contact stack sequence, thickness, additives, and deposition methods.
6. **Extract JV and Stability Data**: Capture JV test results (Voc, Jsc, FF, PCE) and stability measurements (temperature range, atmosphere, PCE degradation).
7. **Handle Missing Data**: If any field cannot be determined, set its value to **”Unknown”** or an empty string ("") as specified in the schema.

Figure 3.5. COT Extraction Prompt

#### *3.4.1.3. Chain-of-Thought (CoT) Prompting*

Chain-of-Thought (CoT) prompting is a more advanced technique that instructs the model to break down the extraction process into sequential steps, mimicking human-like reasoning. Rather than expecting the model to produce structured output in a single inference step, CoT prompts explicitly guide it through intermediate reasoning stages before arriving at the final structured extraction. This approach is particularly beneficial for tasks involving: hierarchical information retrieval where attributes are interconnected (e.g., substrate stack sequences, multi-layer depositions), multi-step extractions where a model needs to extract primary data before deriving secondary attributes, and ambiguous references, where a model needs to correlate different parts of a document before making an inference. Figure 3.5 presents the prompt used for our COT experiment.

### *3.4.2. Fine-tuning*

Fine-tuning is a powerful technique that allows LLMs to specialize in extracting structured information from domain-specific text. While prompt engineering optimizes LLM outputs by guiding inference behavior, fine-tuning directly updates model parameters, making the model inherently more proficient in understanding the variations of PSCs literature.

Beyond enhancing overall extraction proficiency, fine-tuning also addresses several critical challenges associated with domain-specific information retrieval. One of its most significant advantages is domain adaptation. As LLMs are typically pre-trained on broad and diverse corpora, they often lack specialization in niche fields such as PSCs. By fine-tuning on PSCs literature, models can recognize domain-specific terminology, formatting conventions, and understand intricate relationships between extracted attributes, leading to more precise and contextually relevant extractions.

Another significant advantage of fine-tuning is improved consistency in model outputs. It trains the model to extract data in a structured format, following the same patterns present in the fine-tuning data. This reduces inconsistencies and ensures that extracted attributes remain stable across different queries and datasets.

To further enhance the performance of LLMs in extracting structured scientific information, we initially fine-tuned GPT-4o-mini and LLaMA-3.1-70B on a PSCs domain-specific dataset. This fine-tuning process aimed to adapt the models to the language, terminology, format, and structure commonly found in perovskite solar cell literature, thus improving their ability to extract key attributes with higher consistency. However, Qwen will not be fine-tuned due to the computational cost and the additional financial fees.

The training and validation sets from our dataset were used for fine-tuning. After fine-tuning, the test set—which remained unseen during training—was used for evaluating the models' performance, to ensure a fair and unbiased assessment.

## 3.5. Proposed Evaluation Methodology

Evaluating the effectiveness of materials science information extraction is a complex task, particularly when dealing with unstructured data from research papers. To achieve a comprehensive assessment, we employ a rule-based approach designed to systematically evaluate structured attributes, taking into account exact matching, partial matching, and normalization techniques.

Device attributes in PSCs literature are presented in diverse formats, including structured tables, inline numerical values, and descriptive text. This variability necessitates a robust evaluation framework that can accurately assess the correctness of extracted information across various attribute types. To address this challenge, we developed a custom evaluation algorithm that effectively handles both single-device and multi-device cases.

For single-device evaluation, the system compares the extracted attributes against the ground truth when a research paper describes only one device configuration. In contrast, multi-device evaluation is required when multiple distinct device configurations are reported within the same paper, necessitating an advanced matching mechanism to align extracted devices with their corresponding ground-truth records.

### *3.5.1. Different Attribute Types Handling*

Our evaluation algorithm processes the extracted attributes using a schema-based comparison that incorporates various matching techniques, tailored to attribute complexity. Given the diversity of scientific attributes, we implemented a flexible matching strategy that includes:

- Normalization: The first step is done before matching. It standardizes inconsistencies in terminology, abbreviations, and formatting using a predefined equivalence mapping (e.g., mapping ”False” → ”No”, ”TBP” → ”tert-Butylpyridine”, etc.).

- Exact Matching: It's applied to highly structured attributes, such as numerical values, material compositions, and processing parameters, where an exact correspondence is required. For example, if the ground truth is ”Spin-coating” and the prediction is ”Spin-coating”, the prediction is considered correct. However, if the prediction is ”Drop-casting”, it would be marked as incorrect. Similarly, if the ground truth for Perovskite Composition Long Form is ”MAPbI3”, and the prediction is ”MAPbI3”, this would be correct. However, if the predicted value is ”FAPbI3” or ”MAPbI”, it would be incorrect. In the case of numerical attributes, if the ground truth for Perovskite Deposition Thermal Annealing Temperature is 100°C and the prediction is 150°C, it is incorrect.

- Partial Matching: This technique evaluates whether key components of the predicted answer exist in the ground truth, enabling partial credit for predictions that capture some, but not all, of the correct information. A prediction is considered partially correct if the key terms present in the ground truth, even if they include extra context. For instance, if the ground truth is ”Thermal Evaporation” and the prediction is ”deposited by thermal evaporation through a shadow mask”, the prediction would be considered partially correct because it retains part of the information (thermal evaporation) while introducing additional descriptive elements (through a shadow mask). In contrast, an exact match would require the predicted response to be identical to the ground truth without any extra words or modifications.

### *3.5.2. Multi-Device Evaluation*

Extracting information from research papers on PSCs presents a unique challenge due to the presence of multiple devices per publication. Unlike traditional information extraction tasks, where a single entity is extracted per document, PSCs research often describes multiple devices, each with distinct characteristics. This makes direct one-to-one comparisons between extracted and ground-truth devices ineffective, as the number of devices varies across papers.

To address this issue, we employ the Hungarian algorithm [72], [73] for optimal device alignment. The Hungarian algorithm is an optimization algorithm designed to solve the assignment problem in polynomial time. The algorithm ensures that the best-matching extracted device is paired with its corresponding ground truth device, preventing incorrect mappings that could skew evaluation metrics.

To visually represent this process, Figure 3.6 illustrates the application of the Hungarian algorithm for multi-device evaluation. In the figure, one set of nodes represents the predicted devices (d1, d2, d3, etc.), while the other set represents the ground-truth devices (g1, g2, g3, etc.). Edges between these nodes are weighted by the $F_1$ scores, and the Hungarian algorithm seeks the matching that maximizes the sum of the weights ($F_1$ scores) of the edges included in the matching.

Bold lines indicate the best-matched devices, while dashed lines represent other possible matches with lower $F_1$ scores. The figure illustrates various potential scenarios for matching predicted devices with ground truth devices. These cases include scenarios where the number of predicted devices matches the number of ground-truth devices, as well as cases where there are more predicted devices than ground-truth devices or vice versa.

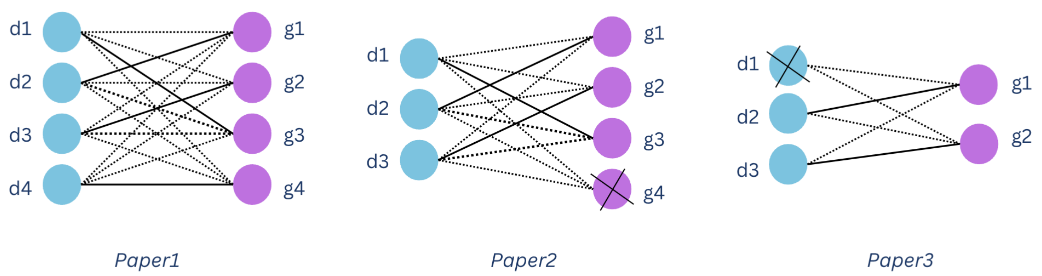


Figure 3.6. Illustration of the Hungarian algorithm applied to multi-device evaluation. Predicted devices (d1, d2, etc.) are matched with ground truth devices (g1, g2, etc.) based on $F_1$ scores. Bold lines indicate the best-matched devices, while dashed lines represent other possible matches with lower $F_1$ scores.

To perform this approach, we followed the following steps:

**Step 1: Attribute-Level Evaluation** Each extracted attribute undergoes a structured comparison using three key matching strategies. For each attribute $a$, we compute precision, recall, and $F_1$-score based on the overlap between the predicted and ground truth attribute components.

The precision and recall for each attribute $a$ in a given device $i$ are computed as follows:

$$\text{Precision}(a,i) = \frac{\text{TP}(a,i)}{\text{TP}(a,i) + \text{FP}(a,i)} \tag{3.1}$$

$$\text{Recall}(a,i) = \frac{\text{TP}(a,i)}{\text{TP}(a,i) + \text{FN}(a,i)} \tag{3.2}$$

where:

- $\text{TP}(a,i)$ (True Positives) counts the correctly predicted components of attribute $a$,
- $\text{FP}(a,i)$ (False Positives) counts predicted components that do not exist in the ground truth, and
- $\text{FN}(a,i)$ (False Negatives) counts ground truth components that were not extracted.

Using these, the $F_1$-score for each attribute $a$ in device $i$ is computed as:

$$F_1(a,i) = \frac{2 \cdot \text{Precision}(a,i) \cdot \text{Recall}(a,i)}{\text{Precision}(a,i) + \text{Recall}(a,i)} \tag{3.3}$$

The overall attribute-level performance is then measured as the mean $F_1$-score across all evaluated attributes:

$$S_{\text{attr}} = \frac{1}{N}\sum_{i=1}^{N} F_1(a,i) \tag{3.4}$$

where $N$ is the total number of attributes evaluated, and $a$ represents an individual attribute.

**Step 2: Pairwise Matching of Extracted vs. Ground Truth Devices:** For each extracted device $d_i$ in a given paper, we compute a pairwise similarity score against every ground truth device $g_j$ in the same paper. The similarity is measured using the $F_1$-score, which balances precision and recall.

The precision and recall for an extracted device $d_i$ and a ground truth device $g_j$ are calculated as:

$$\text{Precision}(d_i, g_j) = \frac{\sum_a \text{TP}(a, i, j)}{\sum_a (\text{TP}(a, i, j) + \text{FP}(a, i, j))} \tag{3.5}$$

$$\text{Recall}(d_i, g_j) = \frac{\sum_a \text{TP}(a, i, j)}{\sum_a (\text{TP}(a, i, j) + \text{FN}(a, i, j))} \tag{3.6}$$

where:

- TP$(a, i, j)$ represents the correctly predicted components for attribute $a$ in device $d_i$ that match the ground truth device $g_j$,
- FP$(a, i, j)$ counts predicted components that do not exist in the ground truth, and
- FN$(a, i, j)$ counts ground truth components that were not extracted.

Using these, the device-level $F_1$-score is computed as:

$$F_1(d_i, g_j) = \frac{2 \cdot \text{Precision}(d_i, g_j) \cdot \text{Recall}(d_i, g_j)}{\text{Precision}(d_i, g_j) + \text{Recall}(d_i, g_j)} \tag{3.7}$$

Since the Hungarian algorithm requires a minimization objective, we define the cost function as the negative $F_1$-score:

$$\text{Cost}(d_i, g_j) = -F_1(d_i, g_j) \tag{3.8}$$

This ensures that the optimization process maximizes the similarity between matched devices by minimizing the total cost.

**Step 3: Applying the Hungarian Algorithm:** The Hungarian algorithm computes the optimal pairing based on the $F_1$ scores, ensuring that each predicted device is matched with a ground truth device that maximizes the $F_1$ score of the pairing. This step is critical as it avoids incorrect device matching, which can lead to skewed evaluation metrics.

To determine the best assignment of predicted devices to ground truth devices, we formulate the optimization problem as follows:

$$A^* = \arg\min_A \sum_{(d_i, g_j) \in A} \text{Cost}(d_i, g_j) \tag{3.9}$$

where $A^*$ represents the optimal assignment of predicted devices $d_i$ to ground truth devices $g_j$. The algorithm selects $A^*$ such that it minimizes the total cost, which is the sum of negative $F_1$ scores, effectively maximizing overall similarity. Each element $(d_i, g_j)$ in $A^*$ is a pairing of a predicted device $d_i$ with a ground truth device $g_j$.

Once the optimal assignment of predicted devices to ground-truth devices is established using the Hungarian algorithm, we address cases where the number of predicted devices differs from the number of ground-truth devices. If the model predicts more devices than exist in the ground truth, the unmatched predicted devices are treated as false positives (FP). Conversely, if the model predicts fewer devices than exist in the ground truth, the missing devices are treated as false negatives (FN) as they correspond to devices that should have been extracted but were not identified by the model.

**Step 4: Paper-Level Multi-Device Evaluation:** Once the optimal device matches are established, we compute paper-level precision, recall, and $F_1$-score based on the sum of the matched device-level $F_1$ scores:

$$P_{\text{paper}} = \frac{\sum_{(d_i,g_j)\in A^*} F_1(d_i, g_j)}{|D|} \tag{3.10}$$

$$R_{\text{paper}} = \frac{\sum_{(d_i,g_j)\in A^*} F_1(d_i, g_j)}{|G|} \tag{3.11}$$

$$F_{1\text{paper}} = \frac{2 \cdot P_{\text{paper}} \cdot R_{\text{paper}}}{P_{\text{paper}} + R_{\text{paper}}} \tag{3.12}$$

where $|D|$ and $|G|$ denote the total number of predicted and ground truth devices, respectively.

### *3.5.3. Best $F_1$ Score*

To better assess the quality of extraction independent of quantity estimation, we also compute the best device $F_1$-score, defined as the highest $F_1$-score among all matched device pairs:

$$F_{1\text{best}} = \max_{(d_i,g_j)\in A^*} F_1(d_i, g_j) \tag{3.13}$$

### *3.5.4. Champion Device Evaluation*

In addition to evaluating all extracted devices, we conduct an evaluation focusing on the champion device— the best-performing device reported in each paper. The champion device is identified based on the highest PCE of their JV reverse scan, selecting the maximum value of reverse scan measurements in the ground truth. For this evaluation, models are prompted to extract only the champion device from the paper. This aims to analyze how well LLMs perform in both single and multiple-device extraction.

# CHAPTER 4: EXPERIMENTAL EVALUATION

In this chapter, we present the experimental setup of our work and discuss the results of the research questions we aim to address.

## 4.1. Experimental Setup

This section details the dataset processing and outlines the setup for prompt engineering, fine-tuning experiments, and evaluation procedures.

### *4.1.1. Data Splitting & Pre-processing*

We divided the dataset based on the number of devices per paper to ensure robust evaluation and generalization. We used a stratified approach to maintain a balanced representation across training, testing, and development sets, dividing the dataset as follows:

- Training set (60%) – Used for fine-tuning experiments.
- Development set (20%) – Used for choosing the best prompt technique.
- Testing set (20%) – Used for evaluating models performance.

This strategy was designed to evenly distribute papers reporting single devices and those reporting multiple devices across all dataset splits, ensuring comprehensive testing and training coverage.

Our dataset comprised documents in both XML and PDF formats. We converted all documents to plain text to facilitate uniformity and compatibility with our models. We utilized PyPDF2[1] library for converting PDF files and ElementTree[2] for parsing and converting XML files. In addition to format conversion, we removed non-essential sections, such as references and acknowledgments, from the documents. The cleaned and converted text files were used as input for our models. We chose not to chunk the papers, as our employed models can handle significant input contexts.

### *4.1.2. Models*

For our experiments, we selected three LLMs: GPT-4o-mini[3], LLaMA-3.1-70B[4], and Qwen-2.5-72B[5]. GPT-4o-mini, developed by OpenAI, is a lightweight version of GPT-4o designed for efficiency while maintaining strong language understanding capabilities. It is accessed through an API and supports a context length of 128K tokens, allowing it to process full research papers without segmentation.

LLaMA-3.1-70B, an open-source model from Meta, supports a 128K token context window, making it suitable for long-form document processing.

Lastly, Qwen 2.5 72B, developed by Alibaba Cloud[6], is optimized for fast inference and scalability. It supports a larger context length of up to 32,768 tokens, making it ideal for processing long scientific documents without truncation.

[1] https://pypi.org/project/PyPDF2/

[2] https://docs.python.org/3/library/xml.etree.elementtree.html

[3] https://openai.com/

[4] https://www.llama.com/

[5] https://www.alibabacloud.com/product/qwen

[6] https://www.alibabacloud.com/product/qwen

Both LLaMA-3.1-70B and Qwen-2.5-72B are large-scale models that require significant computational resources for local deployment. To address this, we accessed these models through the Together.ai API [7], a platform that hosts and manages access to large language models developed by other organizations, such as Meta and Alibaba Cloud. It provides seamless integration and eliminates the need for extensive local infrastructure.

These models were selected based on their architecture, context length, fine-tuning capabilities, and API accessibility, allowing us to explore both prompt-based and fine-tuned approaches for materials science data extraction.

### *4.1.3. Baselines*

To evaluate the effectiveness of our approach, we compare it against two recent SOTA works in automated information extraction from PSCs literature: Valencia et al. [6] and Xie et al. [7]. Since both baselines employ different evaluation methodologies, we re-evaluate our model using the same attribute sets and evaluation metrics reported in these studies.

### *4.1.4. Fine-tuning*

The fine-tuning process was conducted using the respective APIs of the platforms hosting each model. For GPT-4o-mini, we utilized the OpenAI API, which provides direct access to OpenAI's models. For LLaMA 3.1 70B, we employed the Together.ai API. By relying on these APIs, we eliminated the need for local hosting of large models. During fine-tuning, we relied on the default hyperparameters provided by the platform, allowing us to evaluate the models' out-of-the-box adaptability to domain-specific data without additional tuning.

The results of this fine-tuning process are compared against the best-performed prompt from the prompting-based approach to determine the most effective strategy for structured data extraction, as discussed in the following sections.

### *4.1.5. Evaluation Metrics*

To quantitatively assess extraction performance, we compute the following standard metrics:

- Precision: Measures the proportion of correctly extracted attributes out of all extracted attributes.
- Recall: Captures how many of the ground truth attributes were successfully extracted.
- $F_1$ Score: A harmonic mean of precision and recall.
- Best $F_1$ Score: Selects the best $F_1$ score among all predicted devices per paper and then computes the average across all papers, reflecting the model's ability to retrieve an accurate device even when multiple predictions exist.

[7] `https://www.together.ai/`

- Paper-Level $F_1$ Score: Computes the $F_1$ score per paper by aggregating matched device scores using precision and recall at the document level.
- Champion $F_1$ Score: Computes the $F_1$ score of the champion device per paper. Then the score is averaged across all papers to evaluate the model's ability to extract the Champion device from each study.

In addition to these standard evaluation metrics, we introduced an attribute-level analysis, where the $F_1$ score is computed separately for each extracted attribute. For each attribute $a_i$, we compute precision by measuring the proportion of correctly extracted components relative to all extracted components. Recall is calculated by determining how many of the actual ground truth components were successfully extracted. The $F_1$ score for each attribute is then computed as the harmonic mean of precision and recall. This allows us to point out specific attributes that are more challenging to extract and analyze their individual extraction performance across different cases.

## 4.2. Experimental Results

In this work, we aim to answer the following research questions:

**RQ1:** How do different prompt engineering techniques (few-shot, zero-shot, CoT) impact extraction performance?

**RQ2:** How does extracting all devices affect overall extraction performance compared to focusing on one device only?

**RQ3:** Is fine-tuning more effective than extensive prompt engineering for material science data extraction?

**RQ4:** How does our approach compare to SOTA?

### *4.2.1. Effect of Prompt Engineering on Extraction Performance (RQ1)*

To assess the impact of prompt engineering, we compare the performance of zero-shot, few-shot, and CoT prompting across three LLMs, namely GPT-4o-mini, LlaMa-3.1-70B, and Qwen-2.5-72 B. The evaluation is conducted on the development set, assessed through precision (P), recall (R), $F_1$ score, and best device $F_1$ (Best $F_1$) metrics. Table 4.1 shows the performance of each model for all devices vs the champion device experiments presented in table 4.2. The results show that the few-shot approach outperforms other prompting techniques, achieving $F_1$ scores of 0.7735 (Llama) and 0.7687 (Qwen) across all devices, with particularly strong gains in recall (Llama: +25.3% vs zero-shot).

Apparently, all models show improved performance when focusing on champion devices, with Llama achieving $F_1$ of 0.8066 (+17.4% vs all-device performance).

Table 4.1. Impact of Prompt Engineering on Multi-Device Extraction. This table presents the precision (P), recall (R), and $F_1$ scores for Zero-Shot, Few-Shot, and CoT prompting across GPT-4o-mini, LLaMA-3.1-70B, and Qwen-2.5-72 B. The evaluation is conducted on the development set for all devices, using multi-device extraction.

| **Model** | **P** | **R** | **$F_1$** | **Best $F_1$** |
|---|---|---|---|---|
| | | **Zero-Shot** | | |
| GPT | 0.587 | **0.528** | **0.526** | 0.646 |
| Llama | **0.673** | 0.464 | 0.511 | **0.691** |
| Qwen | 0.662 | 0.451 | 0.502 | 0.671 |
| | | **Few-Shot** | | |
| GPT | **0.690** | 0.622 | 0.619 | 0.767 |
| Llama | 0.620 | **0.717** | **0.633** | **0.774** |
| Qwen | 0.619 | 0.701 | 0.624 | 0.769 |
| | | **CoT** | | |
| GPT | 0.615 | **0.522** | **0.537** | 0.653 |
| Llama | 0.671 | 0.475 | 0.517 | **0.697** |
| Qwen | **0.673** | 0.453 | 0.510 | 0.675 |

Table 4.2. Impact of Prompt Engineering on Champion-Device Extraction. This table presents the precision (P), recall (R), and $F_1$ scores for Zero-Shot, Few-Shot, and CoT prompting across GPT-4o-mini, LLaMA-3.1-70B, and Qwen-2.5-72B. The evaluation is conducted on the development set for champion device extraction only.

| **Model** | **P** | **R** | **$F_1$** |
|---|---|---|---|
| | **Zero-Shot** | | |
| GPT | 0.702 | 0.580 | 0.630 |
| Llama | **0.746** | 0.617 | **0.672** |
| Qwen | 0.672 | **0.684** | 0.671 |
| | **Few-Shot** | | |
| GPT | 0.740 | 0.744 | 0.738 |
| Llama | **0.829** | 0.795 | **0.807** |
| Qwen | 0.747 | **0.815** | 0.773 |
| | **CoT** | | |
| GPT | 0.686 | 0.610 | 0.640 |
| Llama | **0.719** | 0.655 | **0.678** |
| Qwen | 0.680 | **0.688** | 0.675 |

The results demonstrate distinct performance across various prompting techniques, highlighting the strengths and trade-offs of each approach. Zero-shot prompting, despite not using examples, achieves moderate precision while maintaining notably good Best $F_1$ scores, indicating that it is capable of extracting well-structured information for at least some devices. In particular, LLaMA achieves a Best $F_1$ of 0.6905 in the multi-device setting. However, recall remains a major limitation, as observed with Qwen (R

= 0.4511) and LLaMA (R = 0.4643), suggesting that without explicit demonstrations, models struggle to consistently capture all relevant details across multiple devices.

On the other hand, few-shot prompting significantly enhances extraction accuracy, offering substantial improvements across all metrics. Providing the model with structured examples allows it to better generalize to device configurations and maintain consistency across extractions. The few-shot approach achieves the highest overall $F_1$ scores, with LLaMA reaching 0.6329 for multi-device extraction. More importantly, the Best $F_1$ scores also improve, reaching 0.7735 for LLaMA, 0.7687 for Qwen, and 0.7667 for GPT, confirming that the alignment between the extracted and ground-truth devices is optimized when examples guide the model's output structure. When focusing on the champion device only, few-shot prompting further enhances extraction accuracy, with LLaMA achieving an $F_1$ of 0.8066, representing a 17.4% increase compared to its multi-device extraction.

As part of our experimentation, we tested one-shot, two-shot, and three-shot prompting to determine the optimal number of examples required for effective extraction. Our findings indicate that three-shot prompting consistently delivered the best performance across all models. As a result, we applied the three-shot setting in all our experiments.

Chain-of-Thought (CoT) prompting, designed to encourage stepwise reasoning, shows performance improvements over zero-shot but does not surpass few-shot prompting. The structured reasoning process enhances recall over zero-shot learning, particularly for models like GPT and LLaMA, which infer missing details by reasoning through multiple steps. However, without explicit examples, CoT still lags behind few-shot prompting in precision and consistency, as models struggle with ambiguous cases where attribute formulations vary. Moreover, CoT prompting performs competitively in terms of Best $F_1$ scores, reaching 0.6970 for LLaMA, which is higher than zero-shot but lower than few-shot, reinforcing that CoT improves attribute alignment but does not fully address recall limitations. This can support the benefit of applying a hybrid approach that combines both COT prompts with few-shot examples.

#### *4.2.2. Impact of Extracting All Devices vs. One Device Only (RQ2)*

To address RQ2, we analyze how evaluation metrics, such as precision, recall, and $F_1$ score, are influenced by the complexity of multi-device reporting. Unlike single-device extraction, where the evaluation focuses on matching attributes for a single ground truth device, multi-device extraction introduces an additional layer of complexity. Here, models should correctly identify multiple devices within a single publication while maintaining high attribute accuracy for each. The challenge lies in ensuring that the extracted devices align with the actual number of devices reported in the research papers, as discrepancies in device count significantly impact evaluation metrics.

We observe that performance variations primarily stem from the alignment between the predicted and actual number of devices reported in papers. A model that extracts an incorrect number of devices, even with high attribute accuracy, may still suffer from penalized recall or precision scores due to mismatches in the expected device count. Given that traditional $F_1$ score calculations depend on both precision (correctly extracted attributes per predicted device) and recall (correctly extracted attributes per ground truth device), discrepancies in device counts negatively influence overall extraction performance.

To better assess the quality of information extraction independent of errors in quantity estimation, we calculated the Best $F_1$ metric, which evaluates the highest $F_1$ score achieved for any single extracted device within a paper and then averages this across the dataset. This approach isolates errors stemming from device count mismatches and focuses on evaluating attribute extraction accuracy independent of quantity estimation errors. Notably, this adjustment reveals significant improvements in performance, such as +18% increase in Best $F_1$ score when using the LLaMA model with the COT and zero-shot experiments, and +14.1% increase in Best $F_1$ score with the few-shot experiment. This suggests that once models correctly recognize the existence of a device, they tend to extract its attributes with reasonable effectiveness.

#### *4.2.2.1. Champion Device*

To better understand the performance of LLMs in extracting a single device per research paper, we introduce a focused evaluation on the Champion Device reported in the study. Unlike the multi-device extraction setup, which requires capturing all devices described in a paper, this evaluation isolates the extraction process to the champion device only.

For this evaluation, the models were prompted to extract only the champion device per paper. The same prompting techniques—zero-shot, few-shot, and Chain-of-Thought (CoT)—were applied, and the fine-tuned GPT model was also assessed under this setting.

As shown in Table 4.2, models exhibit notable improvements in extraction performance when focusing only on the champion device. Compared to the multi-device setting, all models achieve higher $F_1$ scores, with LLaMA reaching 0.807 (+17.4% vs. multi-device).

Few-shot prompting consistently delivers the best results across all models, confirming the importance of in-context examples for structured data extraction. LLaMA (Few-Shot) achieves the highest champion-device $F_1$ score of 0.807, outperforming GPT and Qwen.

Furthermore, fine-tuning demonstrates substantial performance, with GPT-Fine-Tuned achieving $F_1$ of 0.9006, marking a 16.2% improvement over the best-performing prompt.

### *4.2.3. Fine-Tuning vs. Prompt Engineering (RQ3)*

To answer this research question, we examined the effectiveness of domain-specific fine-tuning compared to optimized prompt engineering, observing significant performance enhancements. As shown in Tables 4.3 and 4.4, we present a comparison of the effectiveness of both approaches using the test dataset, where we selected the best-performing prompt from three different prompts and compared its results to those of fine-tuning.

The fine-tuned GPT model demonstrates remarkable $F_1$ scores, achieving 0.9006 for champion devices, and 0.7870 for all devices, which represents a 16% and 17% improvement over the best prompt, respectively. Similarly, the fine-tuned Llama model also demonstrates remarkable performance, with $F_1$ score of 0.8902 for champion devices and 0.7674 for all devices, showing an improvement of 10.8% and 13.5% over the best prompt, respectively.

Furthermore, the Best $F_1$ score of 0.9098 for the fine-tuned GPT model and 0.899 for the fine-tuned Llama model suggest near-optimal extraction performance when devices are accurately identified, indicating that most discrepancies arise from the detection of the ground truth devices rather than from the extraction of their attributes. Notably, 91% of devices correctly identified by GPT and approximately 90% by Llama are extracted with a high accuracy.

Those results demonstrate that fine-tuning (for both open- and closed-source LLMs) is the most effective approach for extracting materials science information, while also highlighting the need for an automated extraction pipeline that minimizes human-induced errors and provides consistent, scalable data extraction for PSCs research.

Table 4.3. Comparison of Fine-Tuning vs. Prompt Engineering for Multi-Device Extraction. The best-performing prompt (from zero-shot, few-shot, and chain-of-thought) is reported for each model. Fine-tuning results are provided for GPT on all extracted devices.

| **Model** | **P** | **R** | **$F_1$** | **Best $F_1$** |
|---|---|---|---|---|
| | | **All Devices** | | |
| GPT | 0.609 | **0.711** | 0.618 | 0.766 |
| Llama | **0.683** | 0.648 | **0.631** | 0.767 |
| Qwen | 0.602 | 0.707 | 0.613 | 0.770 |
| | | **Fine-Tuning** | | |
| GPT | **0.817** | **0.828** | **0.787** | **0.910** |
| Llama | 0.811 | 0.799 | 0.7674 | 0.899 |

Table 4.4. Comparison of Fine-Tuning vs. Prompt Engineering for Champion Device Extraction. The best-performing prompt (from zero-shot, few-shot, and chain-of-thought) is reported for each model. Fine-tuning results are provided for GPT on champion device extraction.

| **Model** | **P** | **R** | **$F_1$** |
|---|---|---|---|
| | **Champion Device** | | |
| GPT | 0.7501 | 0.7400 | 0.7392 |
| Llama | **0.8329** | 0.7868 | **0.8020** |
| Qwen | 0.7472 | **0.8175** | 0.7753 |
| | **Fine-Tuning** | | |
| GPT | 0.9067 | **0.8988** | **0.9006** |
| Llama | **0.9177** | 0.8699 | 0.8902 |

#### *4.2.3.1. Attribute-level Evaluation*

To gain deeper insights into the effectiveness of the fine-tuned model, we evaluated its extraction performance at the attribute level. This analysis allows us to identify which attributes are extracted with high accuracy and where the model tends to fail, providing valuable information for future enhancements. To assess the effectiveness of fine-tuning for each attribute, we computed the $F_1$ score per attribute, measuring precision and recall variations across different extracted fields. All results reported in this sub-section are based on the outputs of the GPT fine-tuned model.

As illustrated in Figures 4.1 and 4.2, the majority of attributes achieve high $F_1$ scores, demonstrating the model's effectiveness in structured data extraction. Notably, several attributes, such as ”Cell_semitransparent_wavelength_range” and ”Perovskite_surface_treatment_before_next_deposition_step”, achieve perfect or near-perfect extraction ($F_1$ = 1.0).

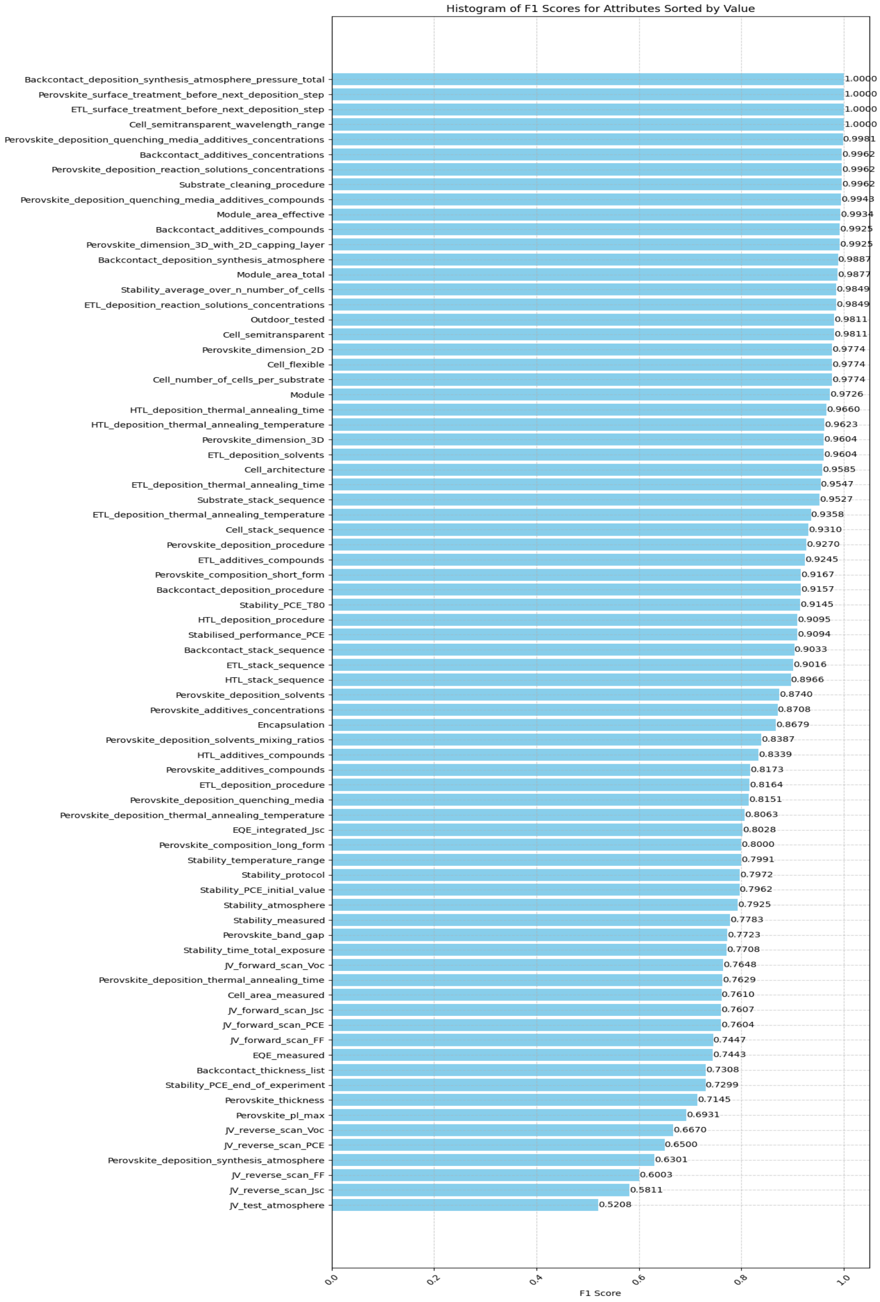


Figure 4.1. Per-attribute $F_1$ score analysis of the fine-tuned GPT. The histogram presents the $F_1$ scores for each extracted attribute, highlighting variations in model performance across different attribute categories.

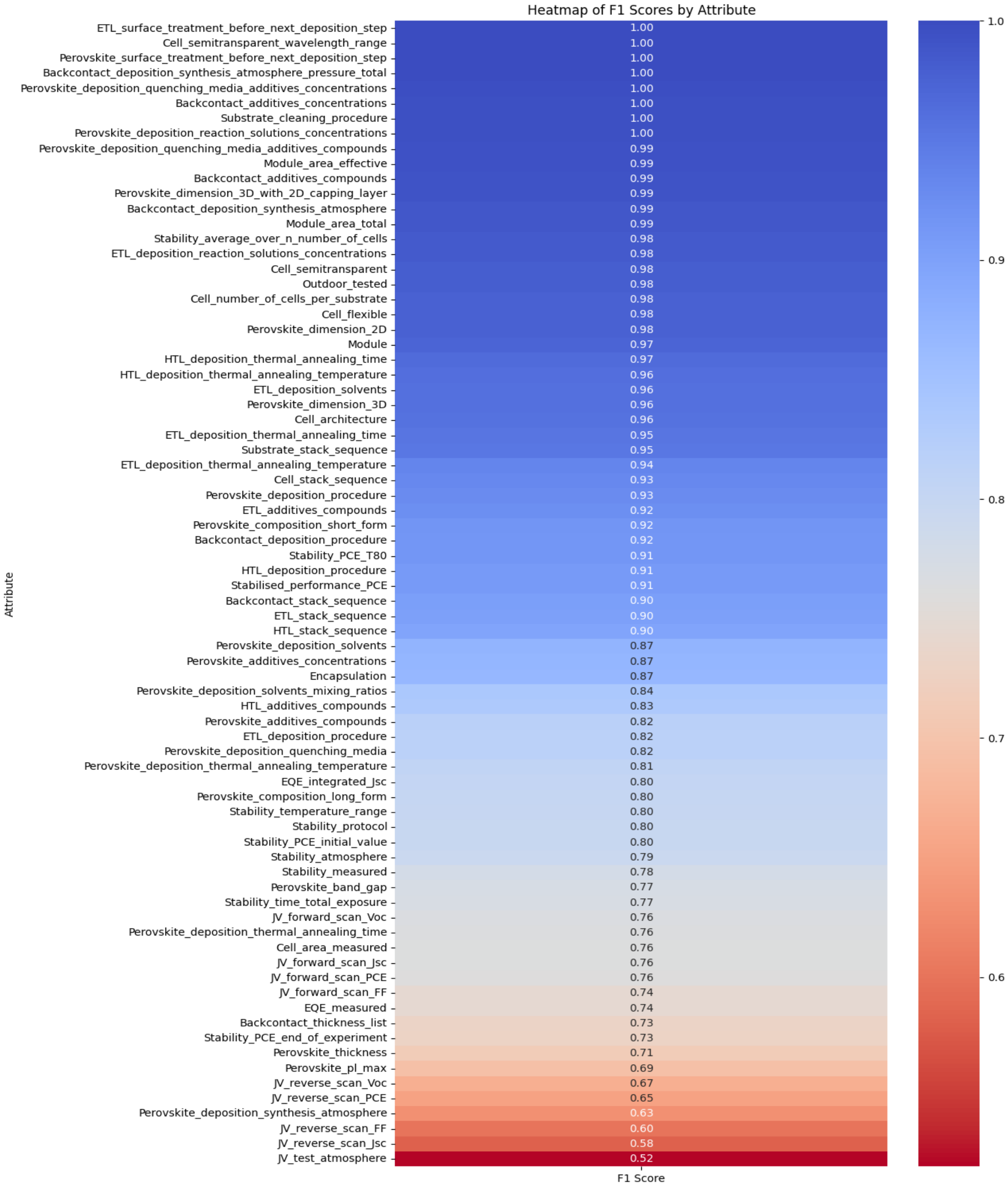


Figure 4.2. Heatmap representation of $F_1$ score for each extracted attribute in PSCs literature. The color intensity reflects extraction performance, with higher scores indicating better attribute extraction accuracy.

On the other hand, some attributes exhibit moderate to lower extraction performance. Among the least accurately extracted attributes are ”JV_test_atmosphere” ($F_1$ = 0.5208), ”JV_reverse_scan_Jsc” ($F_1$ = 0.5811), and ”Perovskite_deposition_synthesis_atmosphere” ($F_1$ = 0.6301).

Additionally, attributes related to electrical performance metrics (e.g., ”JV_reverse_scan_PCE”, ”JV_reverse_scan_Voc”, and ”EQE_measured”) exhibit relatively lower $F_1$ scores compared to material composition and structural attributes. Interestingly, material composition-related attributes (e.g., ”Perovskite_composition_long_-

form", "Perovskite_additives_compounds", and "Perovskite_deposition_thermal_annealing_temperature").

Overall, these findings highlight areas where the model performs exceptionally well and areas where further refinements are necessary. Improving the extraction of performance-related attributes and processing conditions could enhance the completeness and usability of extracted data for scientific analysis. Future work may explore techniques such as schema adjustments, hybrid rule-based post-processing, or additional fine-tuning with accurate and correct annotations to enhance the accuracy of attribute-level extraction further.

### *4.2.4. Comparison with SOTA (RQ4)*

To objectively evaluate the effectiveness of our approach, we conduct a comparison with two recent state-of-the-art (SOTA) works in PSCs information extraction: [6] and [7].

Both SOTA works focus only on single-device extraction and do not account for multi-device cases. In contrast, our approach is designed to handle multi-device scenarios by extracting and evaluating all reported devices from each paper. To ensure a fair comparison with the SOTA models, we restrict this evaluation to the champion-device extraction results from the fine-tuned GPT model.

The set of attributes used in [7] is entirely contained within our selection, making it a direct subset. In contrast, the attributes chosen in [6] are also included in our dataset (except for one attribute, namely *"HTL_annealing_parameters"*) but are structured differently. Their approach decomposes specific attributes into multiple separate fields, whereas our schema consolidates them into single unified fields. A key example of this structural difference is the representation of perovskite composition. In [6], this attribute is split across several separate fields:

> "Perovskite_composition_short_form, Perovskite_composition_a_ions, Perovskite_composition_a_ions_coefficients, Perovskite_composition_b_ions, Perovskite_composition_b_ions_coefficients, Perovskite_composition_c_ions, and Perovskite_composition_c_ions_coefficients".

In our schema, all this information is captured within a single attribute, namely, *"Perovskite_composition_long_form"*.

We developed a custom code that automatically decomposes our unified attributes into the equivalent attributes used in [6]. This step enables the direct alignment of extracted values between approaches, thereby eliminating structural discrepancies that could otherwise skew the evaluation. Table 4.7 provides a visual representation of the attribute selections across all three approaches.

Although the work of Valencia et al. [6] employs an attribute-level evaluation methodology, where per-attribute scores are averaged to assess extraction performance, this approach still differs in dataset curation, annotation schema, and attribute definitions, preventing a direct one-to-one comparison.

On the other hand, the work of [7] computes evaluation metrics using an aggregated approach, where precision, recall, and $F_1$ scores are calculated based on the total number of correctly extracted values across all attributes. This leads to a different distribution of evaluation results, making direct comparison unfair. We recalculated our evaluation metrics using the same aggregated approach employed by [7]. As shown in Table 4.5, our approach outperforms the method proposed by [7] across all metrics, achieving higher precision, recall, and $F_1$.

Table 4.5. Comparison of our approach with Xie et al. [7]. The evaluation follows the aggregate-based computation and the same set of attributes used in their work.

| **Attributes** | **Our Work** | | | **Xie et al. [7]** | | |
|---|---|---|---|---|---|---|
| | **P** | **R** | $\boldsymbol{F_1}$ | **P** | **R** | $\boldsymbol{F_1}$ |
| Xie et al. Attributes | **0.906** | **0.899** | **0.902** | 0.883 | 0.861 | 0.871 |

In addition to this comparison, we also highlight a key contrast in the effectiveness of direct prompting techniques. In [7], the authors reported that direct prompting using GPT-3.5 yielded poor performance, particularly for the IE task, where an $F_1$ score of only 28.7% was achieved. They concluded that direct prompting was unreliable for structured data extraction and suggested that fine-tuning was necessary for obtaining meaningful results. Table 4.6 summarizes this comparison.

Table 4.6. Comparison of state-of-the-art direct prompting results with our optimized prompt engineering techniques. The evaluation follows the aggregate-based computations used in the baseline work. This experiment was conducted on the development set.

| **Model** | **Prompting Technique** | **P** | **R** | $\boldsymbol{F_1}$ |
|---|---|---|---|---|
| GPT-3.5 [7] | Direct Prompting | 0.226 | 0.430 | 0.287 |
| GPT | Zero-Shot | 0.738 | 0.499 | 0.595 |
| | Few-Shot | 0.782 | 0.703 | 0.740 |
| | CoT | 0.720 | 0.558 | 0.629 |
| Llama | Zero-Shot | 0.719 | 0.579 | 0.641 |
| | Few-Shot | **0.848** | 0.761 | **0.802** |
| | CoT | 0.742 | 0.554 | 0.635 |
| Qwen | Zero-Shot | 0.701 | 0.627 | 0.662 |
| | Few-Shot | 0.804 | **0.785** | 0.795 |
| | CoT | 0.701 | 0.630 | 0.663 |

However, our results challenge this conclusion. By carefully designing effective prompts, we demonstrate that direct prompting can achieve significantly better results. Specifically, with a simple 3-shot prompting approach, we obtained an $F_1$ score of 80% for the champion device extraction. Moreover, while few-shot prompting achieves the highest results, even zero-shot and CoT prompting demonstrate exemplary performance. Notably, Qwen achieves around 66% in both COT and zero-shot, while LLaMA achieves 64.14% in zero-shot and 63.46% in CoT prompting.

In future work, we plan to apply the evaluation methodologies used by both SOTA approaches to our dataset. By re-evaluating our extraction pipeline under their specific evaluation frameworks, we can directly assess how our approach performs within their constraints.

Rather than comparing raw numbers only, we additionally focus on methodological advancements that differentiate our approach:

- Multi-Device Extraction: Unlike prior works that primarily extract a single device per paper, our model is designed to handle multiple devices while maintaining per-device accuracy.
- Schema Generalization: Our schema encompasses a broader range of attributes, providing a more comprehensive representation of PSC literature.
- Prompt Engineering vs. Fine-Tuning: While previous studies emphasize fine-tuning, we demonstrate that optimized prompting can achieve competitive performance, offering a more scalable approach.

By emphasizing these methodological aspects rather than direct performance numbers, we provide a more balanced and informative comparison, avoiding misleading interpretations due to dataset discrepancies. Future work should explore benchmarking under a unified dataset.

Table 4.7. Attributes comparison with state-of-the-art works. The symbol ✓ denotes supported attributes, while × indicates attributes that are not supported.

| Attribute Name | Our Work | Valencia et al. [6] | Xie et al. [7] |
|---|---|---|---|
| Ref_DOI_number | ✓ | ✓ | ✓ |
| Cell_stack_sequence | ✓ | × | × |
| Cell_area_measured | ✓ | × | ✓ |
| Cell_number_of_cells_per_substrate | ✓ | × | ✓ |
| Cell_architecture | ✓ | ✓ | ✓ |
| Cell_flexible | ✓ | ✓ | ✓ |
| Cell_semitransparent | ✓ | × | ✓ |
| Cell_semitransparent_wavelength_range | ✓ | × | ✓ |
| Module | ✓ | × | ✓ |
| Module_area_total | ✓ | × | × |
| Module_area_effective | ✓ | × | × |
| Substrate_stack_sequence | ✓ | ✓ | ✓ |
| Substrate_cleaning_procedure | ✓ | × | × |
| ETL_stack_sequence | ✓ | × | ✓ |
| ETL_additives_compounds | ✓ | × | ✓ |
| ETL_deposition_procedure | ✓ | ✓ | ✓ |
| ETL_deposition_solvents | ✓ | × | × |
| ETL_deposition_reaction_solutions_concentrations | ✓ | × | × |
| ETL_deposition_thermal_annealing_temperature | ✓ | ✓ | ✓ |
| ETL_deposition_thermal_annealing_time | ✓ | ✓ | ✓ |
| ETL_surface_treatment_before_next_deposition_step | ✓ | × | × |
| Perovskite_dimension_2D | ✓ | × | × |
| Perovskite_dimension_3D | ✓ | × | × |
| Perovskite_dimension_3D_with_2D_capping_layer | ✓ | × | × |
| Perovskite_composition_long_form | ✓ | × | ✓ |
| Perovskite_additives_compounds | ✓ | × | ✓ |
| Perovskite_additives_concentrations | ✓ | × | × |
| Perovskite_thickness | ✓ | × | × |
| Perovskite_band_gap | ✓ | × | × |
| Perovskite_pl_max | ✓ | × | × |
| Perovskite_deposition_procedure | ✓ | ✓ | ✓ |
| Perovskite_deposition_synthesis_atmosphere | ✓ | × | ✓ |
| Perovskite_deposition_solvents | ✓ | × | ✓ |
| Perovskite_deposition_solvents_mixing_ratios | ✓ | × | × |
| Perovskite_deposition_reaction_solutions_concentrations | ✓ | × | × |
| Perovskite_deposition_quenching_media | ✓ | × | × |
| Perovskite_deposition_quenching_media_additives_compounds | ✓ | × | × |
| Perovskite_deposition_quenching_media_additives_concentrations | ✓ | × | × |
| Perovskite_deposition_thermal_annealing_temperature | ✓ | ✓ | ✓ |
| Perovskite_deposition_thermal_annealing_time | ✓ | ✓ | ✓ |
| Perovskite_surface_treatment_before_next_deposition_step | ✓ | × | × |
| HTL_stack_sequence | ✓ | × | ✓ |
| HTL_deposition_procedure | ✓ | ✓ | ✓ |
| HTL_deposition_thermal_annealing_temperature | ✓ | ✓ | × |
| HTL_deposition_thermal_annealing_time | ✓ | ✓ | × |
| Backcontact_stack_sequence | ✓ | × | ✓ |
| Backcontact_thickness_list | ✓ | × | × |
| Backcontact_additives_concentrations | ✓ | × | × |
| Backcontact_deposition_synthesis_atmosphere | ✓ | × | ✓ |
| Backcontact_deposition_synthesis_atmosphere_pressure_total | ✓ | × | × |
| Encapsulation | ✓ | × | × |
| JV_test_atmosphere | ✓ | × | × |
| JV_reverse_scan_Voc | ✓ | × | × |
| JV_reverse_scan_Jsc | ✓ | × | × |
| JV_reverse_scan_FF | ✓ | × | × |
| JV_reverse_scan_PCE | ✓ | × | × |
| JV_forward_scan_Voc | ✓ | × | × |
| JV_forward_scan_Jsc | ✓ | × | × |
| JV_forward_scan_FF | ✓ | × | × |
| JV_forward_scan_PCE | ✓ | × | × |
| Stabilised_performance_PCE | ✓ | × | × |
| EQE_measured | ✓ | × | × |
| EQE_integrated_Jsc | ✓ | × | × |
| Stability_measured | ✓ | × | ✓ |
| Stability_protocol | ✓ | × | ✓ |
| Stability_temperature_range | ✓ | × | ✓ |

Continued on next page

| **Attribute Name** | **Our Work** | **Valencia et al. [6]** | **Xie et al. [7]** |
|---|---|---|---|
| Stability_atmosphere | ✓ | × | ✓ |
| Stability_time_total_exposure | ✓ | × | ✓ |
| Stability_PCE_initial_value | ✓ | × | ✓ |
| Stability_PCE_end_of_experiment | ✓ | × | ✓ |
| Stability_PCE_T80 | ✓ | × | × |
| Outdoor_tested | ✓ | × | × |
| Stability_average_over_n_number_of_cells | ✓ | × | ✓ |
| Perovskite_composition_short_form | ✓ | ✓ | ✓ |
| HTL_additives_compounds | ✓ | × | ✓ |
| Backcontact_additives_compounds | ✓ | × | ✓ |
| Backcontact_deposition_procedure | ✓ | × | ✓ |

# CHAPTER 5: CONCLUSION AND FUTURE WORK

## 5.1. Conclusions

In this thesis, we presented an automated information extraction pipeline for PSCs research, addressing the challenges posed by multi-device reporting in scientific literature. Our approach leverages LLMs with prompt engineering and fine-tuning to extract structured data from PSCs studies. We introduce the first LLM-based approach for extracting data from multiple devices per publication, ensuring that each device configuration is accurately represented. Furthermore, we are the first to develop an evaluation framework capable of assessing both single-device and multi-device extractions. This framework effectively aligns extracted devices with their ground truth, overcoming the many-to-many matching problem in PSCs research. Additionally, we carefully selected and defined a broad set of attributes most relevant to PSCs researchers, ensuring that the extracted data is both scientifically valuable and practically applicable.

Our experiments demonstrate that the proposed pipeline outperforms existing approaches, providing a scalable, reliable, and structured solution for automated data extraction and evaluation. The results demonstrate the effectiveness of the proposed pipeline, achieving $F_1$ score of 90.06% for champion-device extraction, $F_1$ score of 78.70% for multi-device extraction, and a best $F_1$ score of 90.98% for the best extracted device in multi-device cases. Fine-tuning significantly enhances extraction performance, making LLMs more specialized and accurate for PSCs data. At the same time, prompt engineering proved to be a strong alternative approach, achieving good results through carefully designed prompts. This work lays a strong foundation for automated structured data extraction in materials science, with potential applications extending beyond PSCs to other domains in the field.

## 5.2. Future Work

Despite significant progress in automating information extraction and evaluation from PSCs literature, several areas can be further enhanced. One promising approach is to improve the evaluation framework by incorporating LLM-based evaluators that assess extraction correctness through semantic reasoning. This enhancement would reduce reliance on manual annotations and enhance the robustness of evaluation by considering contextual and implicit relationships within the extracted attributes.

Additionally, scientific publications often present critical information not only in textual descriptions but also in figures, tables, and equations. Traditional text-based extraction approaches may overlook valuable data embedded in these non-textual elements. A future direction involves integrating multimodal extraction techniques to extract structured data from both text and visual components. This could improve extraction completeness and accuracy, particularly for attributes reported exclusively in graphical representations or tabular formats.

Moreover, in this work, we limited our extraction to papers that contain up to four devices. Expanding the threshold to include publications with a higher number of devices could further validate the scalability of our approach and enhance dataset coverage. Additionally, many PSCs studies report essential experimental details in the supplementary information rather than in the main text. Including supplementary information in the dataset would provide a more holistic extraction, capturing missing details and improving overall data completeness.

Finally, while our methodology has been designed explicitly for PSCs research, the principles and techniques developed in this work can be extended to other domains in materials science.

# REFERENCES

[1] K. A. *et al.*, "The 2019 materials by design roadmap," en, *Journal of Physics D: Applied Physics*, vol. 52, p. 013 001, 2018. DOI: 10.1088/1361-6463/aad926.

[2] R. J. *et al.*, "Materials 4.0: Materials big data enabled materials discovery," en, *Applied Materials Today*, vol. 10, pp. 127–132, 2018. DOI: 10.1016/j.apmt.2017.12.015.

[3] C. W. C. *et al.*, "A robotic platform for flow synthesis of organic compounds informed by AI planning," en, *Science*, vol. 365, p. 1566, 2019. DOI: 10.1126/science.aax1566.

[4] R. G.-B. *et al.*, "Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach," en, *Nature Materials*, vol. 15, p. 1120, 2016. DOI: 10.1038/nmat4717.

[5] T. J. Jacobsson, A. Hultqvist, and et al., "An open-access database and analysis tool for perovskite solar cells based on the fair data principles," *Nature Energy*, vol. 7, no. 1, pp. 107–115, 2022.

[6] A. Valencia, F. Liu, X. Zhang, X. Bo, W. Li, and W. A. Daoud, "Auto-generating a database on the fabrication details of perovskite solar devices," *Scientific Data*, vol. 12, no. 1, p. 270, 2025.

[7] T. Xie, Y. Wan, Y. Zhou, *et al.*, "Creation of a structured solar cell material dataset and performance prediction using large language models," *Patterns*, vol. 5, no. 5, 2024.

[8] J. M. Beard Edward J. Cole, "Perovskite- and dye-sensitized solar-cell device databases auto-generated using chemdataextractor," *Scientific Data*, vol. 9, no. 1, 2022. DOI: 10.1038/s41597-022-01355-w.

[9] F. H. *et al.*, "Next-generation experimentation with self-driving laboratories," en, *Trends in Chemistry*, vol. 1, pp. 282–291, 2019. DOI: 10.1016/j.trechm.2019.02.007.

[10] S. Curtarolo, W. Setyawan, G. L. Hart, *et al.*, "Aflow: An automatic framework for high-throughput materials discovery," *Computational Materials Science*, vol. 58, pp. 218–226, 2012, ISSN: 0927-0256. DOI: https://doi.org/10.1016/j.commatsci.2012.02.005. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0927025612000717.

[11] J. E. Saal, S. Kirklin, M. Aykol, *et al.*, "Materials design and discovery with high-throughput density functional theory: The open quantum materials database (oqmd)," *JOM*, vol. 65, no. 11, pp. 1501–1509, Nov. 2013. DOI: 10.1007/s11837-013-0755-4. [Online]. Available: https://doi.org/10.1007/s11837-013-0755-4.

[12] S. Haastrup, M. Strange, M. Pandey, *et al.*, "The computational 2d materials database: High-throughput modeling and discovery of atomically thin crystals," *2D Materials*, vol. 5, no. 4, p. 042 002, Sep. 2018. DOI: 10.1088/2053-1583/aacfc1. [Online]. Available: https://dx.doi.org/10.1088/2053-1583/aacfc1.

[13] C. Draxl and M. Scheffler, "The nomad laboratory: From data sharing to artificial intelligence," *Journal of Physics: Materials*, vol. 2, no. 3, p. 036 001, May 2019. DOI: 10.1088/2515-7639/ab13bb. [Online]. Available: https://dx.doi.org/10.1088/2515-7639/ab13bb.

[14] K. T. Winther, M. J. Hoffmann, J. R. Boes, O. Mamun, M. Bajdich, and T. Bligaard, "Materials design and discovery with high-throughput density functional theory: The open quantum materials database (oqmd)," *JOM*, vol. 6, no. 1, p. 75, May 2019. DOI: 10.1038/s41597-019-0081-y. [Online]. Available: https://doi.org/10.1038/s41597-019-0081-y.

[15] M. Alvarez-Moreno, C. de Graaf, N. López, F. Maseras, J. M. Poblet, and C. Bo, "Managing the computational chemistry big data problem: The iochem-bd platform," *Journal of Chemical Information and Modeling*, vol. 55, no. 1, pp. 95–103, 2015, PMID: 25469626. DOI: 10.1021/ci500593j. [Online]. Available: https://doi.org/10.1021/ci500593j.

[16] G. Bergerhoff, R. Hundt, R. Sievers, and I. D. Brown, "The inorganic crystal structure database," *Journal of Chemical Information and Computer Sciences*, vol. 23, no. 2, pp. 66–69, 1983. DOI: \url{10.1021/ci00038a003}. eprint: https://doi.org/10.1021/ci00038a003. [Online]. Available: https://doi.org/10.1021/ci00038a003.

[17] S. Grazulis, D. Chateigner, R. T. Downs, *et al.*, "Crystallography Open Database – an open-access collection of crystal structures," *Journal of Applied Crystallography*, vol. 42, no. 4, pp. 726–729, Aug. 2009. DOI: 10.1107/S0021889809016690. [Online]. Available: %5Curl%7Bhttps://doi.org/10.1107/S0021889809016690%7D.

[18] J. H. Lee, M. Lee, and K. Min, "Natural language processing techniques for advancing materials discovery: A short review," *International Journal of Precision Engineering and Manufacturing-Green Technology*, vol. 10, no. 5, pp. 1337–1349, 2023.

[19] M. C. Swain and J. M. Cole, "Chemdataextractor: A toolkit for automated extraction of chemical information from the scientific literature," *Journal of chemical information and modeling*, vol. 56, no. 10, pp. 1894–1904, 2016.

[20] Q. Deng and B. Lin, "Automated machine learning structure-composition-property relationships of perovskite materials for energy conversion and storage," *Energy Mater*, vol. 1, p. 100 006, 2021.

[21] T. Brown, B. Mann, N. Ryder, *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[22] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[23] J. Dagdelen, A. Dunn, S. Lee, *et al.*, "Structured information extraction from scientific text with large language models," *Nature Communications*, vol. 15, no. 1, p. 1418, 2024.

[24] K. M. Jablonka, Q. Ai, A. Al-Feghali, *et al.*, "14 examples of how llms can transform materials science and chemistry: A reflection on a large language model hackathon," *Digital discovery*, vol. 2, no. 5, pp. 1233–1250, 2023.

[25] S. Miret and N. M. Krishnan, "Are llms ready for real-world materials discovery?" *arXiv preprint arXiv:2402.05200*, 2024.

[26] X. Wang, L. Huang, S. Xu, and K. Lu, “How does a generative large language model perform on domain-specific information extraction? a comparison between gpt4 and a rule-based method on band gap extraction,” *Journal of Chemical Information and Modeling*, vol. 64, no. 20, pp. 7895–7904, 2024.

[27] M. Schilling-Wilhelmi, M. Ríos-García, S. Shabih, *et al.*, “From text to insight: Large language models for materials science data extraction,” *arXiv preprint arXiv:2407.16867*, 2024.

[28] T. Xie, Y. Wan, W. Huang, *et al.*, “Large language models as master key: Unlocking the secrets of materials science with gpt,” *arXiv preprint arXiv:2304.02213*, 2023.

[29] J. M. F. *et al.*, “Atomistic origins of high-performance in hybrid halide perovskite solar cells,” en, *Nano Letters*, vol. 14, pp. 2584–2590, 2014. DOI: 10.1021/nl500390f.

[30] A. K. *et al.*, “Organometal halide perovskites as visible-light sensitizers for photovoltaic cells,” en, *Journal of the American Chemical Society*, vol. 131, pp. 6050–6051, 2009. DOI: 10.1021/ja809598r.

[31] M. M. L. *et al.*, “Efficient hybrid solar cells based on meso-superstructured organometal halide perovskites,” en, *Science*, vol. 338, pp. 643–647, 2012. DOI: 10.1126/science.1228604.

[32] M. M. B. *et al.*, “Field-tunable quantum disordered ground state in the triangular-lattice antiferromagnet $NaYbO_2$,” en, *Nature Physics*, vol. 15, pp. 1058–1064, 2019. DOI: 10.1038/s41567-019-0594-5.

[33] J.-H. I. *et al.*, “6.5% efficient perovskite quantum-dot-sensitized solar cell,” it, *Nanoscale*, vol. 3, pp. 4088–4093, 2011. DOI: 10.1039/C1NR10867K.

[34] H.-S. K. *et al.*, “Lead iodide perovskite sensitized all-solid-state submicron thin film mesoscopic solar cell with efficiency exceeding 9%,” en, *Scientific Reports*, vol. 2, p. 591, 2012. DOI: 10.1038/srep00591.

[35] K. A. B. *et al.*, “23.6%-efficient monolithic perovskite/silicon tandem solar cells with improved stability,” en, *Nature Energy*, vol. 2, p. 17 009, 2017. DOI: 10.1038/nenergy.2017.9.

[36] Z.-K. T. *et al.*, “Bright light-emitting diodes based on organometal halide perovskite,” en, *Nat Nano*, vol. 9, pp. 687–692, 2014. DOI: 10.1038/nnano.2014.149.

[37] H. C. *et al.*, “Overcoming the electroluminescence efficiency limitations of perovskite light-emitting diodes,” en, *Science*, vol. 350, pp. 1222–1225, 2015. DOI: 10.1126/science.aad1818.

[38] M. Y. *et al.*, “Perovskite energy funnels for efficient light-emitting diodes,” it, *Nat Nano*, vol. 11, pp. 872–877, 2016. DOI: 10.1038/nnano.2016.110.

[39] G. L. *et al.*, “Highly efficient perovskite nanocrystal light-emitting diodes enabled by a universal crosslinking method,” en, *Advanced Materials*, vol. 28, pp. 3528–3534, 2016. DOI: 10.1002/adma.201600064.

[40] F. Z. *et al.*, “Brightly luminescent and color-tunable colloidal $CH_3NH_3PbX_3$ (x = Br, i, Cl) quantum dots: Potential alternatives for display technology,” en, *ACS Nano*, 2015. DOI: 10.1021/acsnano.5b01154.

[41] S. D. S. *et al.*, “Electron-hole diffusion lengths exceeding 1 micrometer in an organometal trihalide perovskite absorber,” en, *Science*, vol. 342, pp. 341–344, 2013. DOI: 10.1126/science.1243982.

[42] Y. H. *et al.*, “Facile synthesis of stable and highly luminescent methylammonium lead halide nanocrystals for efficient light emitting devices,” en, *Journal of the American Chemical Society*, vol. 141, pp. 1269–1279, 2019. DOI: 10.1021/jacs.8b09706.

[43] S. G. M. *et al.*, “Controlling competing photochemical reactions stabilizes perovskite solar cells,” en, *Nature Photonics*, vol. 13, pp. 532–539, 2019. DOI: 10.1038/s41566-019-0435-1.

[44] J. Mavracic, C. J. Court, T. Isazawa, S. R. Elliott, and J. M. Cole, “Chemdataextractor 2.0: Autopopulated ontologies for materials science,” *Journal of Chemical Information and Modeling*, vol. 61, no. 9, pp. 4280–4289, 2021.

[45] H. Huo, Z. Rong, O. Kononova, *et al.*, “Semi-supervised machine-learning classification of materials synthesis procedures,” *Npj Computational Materials*, vol. 5, no. 1, p. 62, 2019.

[46] J. Wei, X. Chu, X.-Y. Sun, *et al.*, “Machine learning in materials science,” *InfoMat*, vol. 1, no. 3, pp. 338–358, 2019.

[47] A. Trewartha, N. Walker, H. Huo, *et al.*, “Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science,” *Patterns*, vol. 3, no. 4, 2022.

[48] I. Beltagy, K. Lo, and A. Cohan, “Scibert: A pretrained language model for scientific text,” *arXiv preprint arXiv:1903.10676*, 2019.

[49] T. Gupta, M. Zaki, N. A. Krishnan, and Mausam, “Matscibert: A materials domain language model for text mining and information extraction,” *npj Computational Materials*, vol. 8, no. 1, p. 102, 2022.

[50] S. Huang and J. M. Cole, “Batterybert: A pretrained language model for battery database enhancement,” *Journal of chemical information and modeling*, vol. 62, no. 24, pp. 6365–6377, 2022.

[51] J. Zhao, S. Huang, and J. M. Cole, “Opticalbert and opticaltable-sqa: Text-and table-based language models for the optical-materials domain,” *Journal of Chemical Information and Modeling*, vol. 63, no. 7, pp. 1961–1981, 2023.

[52] Y. Song, S. Miret, and B. Liu, “Matsci-nlp: Evaluating scientific language models on materials science language tasks using text-to-schema modeling,” in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, pp. 3621–3639.

[53] J. Choi and B. Lee, “Accelerating materials language processing with large language models,” *Communications Materials*, vol. 5, no. 1, p. 13, 2024.

[54] K. Hatakeyama-Sato, N. Yamane, Y. Igarashi, Y. Nabae, and T. Hayakawa, “Prompt engineering of gpt-4 for chemical research: What can/cannot be done?” *Science and Technology of Advanced Materials: Methods*, vol. 3, no. 1, p. 2 260 300, 2023.

[55] D. Zhang, W. Liu, Q. Tan, *et al.*, “Chemllm: A chemical large language model,” *arXiv preprint arXiv:2402.06852*, 2024.

[56] T. Xie, Y. Wan, W. Huang, *et al.*, "Darwin series: Domain specific large language models for natural science," *arXiv preprint arXiv:2308.13565*, 2023.

[57] Z.-Y. Chen, F.-K. Xie, M. Wan, *et al.*, "Matchat: A large language model and application service platform for materials science," *Chinese Physics B*, vol. 32, no. 11, p. 118 104, 2023.

[58] C. Chen, A. Maqsood, Z. Zhang, *et al.*, "The use of chatgpt to generate experimentally testable hypotheses for improving the surface passivation of perovskite solar cells," *Cell Reports Physical Science*, vol. 5, no. 7, 2024.

[59] Y. J. Park, D. Kaplan, Z. Ren, *et al.*, "Can chatgpt be used to generate scientific hypotheses?" *Journal of Materiomics*, vol. 10, no. 3, pp. 578–584, 2024.

[60] L. Zhang, Y. Huang, L. Yan, *et al.*, "Fast exploring literature by language machine learning for perovskite solar cell materials design," *Advanced Intelligent Systems*, p. 2 300 678, 2024.

[61] S. Gupta, A. Mahmood, P. Shetty, A. Adeboye, and R. Ramprasad, "Data extraction from polymer literature using large language models," *Communications Materials*, vol. 5, no. 1, p. 269, 2024.

[62] M. P. Polak and D. Morgan, "Extracting accurate materials data from research papers with conversational language models and prompt engineering," *Nature Communications*, vol. 15, no. 1, p. 1569, 2024.

[63] S. X. Leong, S. Pablo-García, Z. Zhang, and A. Aspuru-Guzik, "Automated electrosynthesis reaction mining with multimodal large language models (mllms)," *Chemical Science*, vol. 15, no. 43, pp. 17 881–17 891, 2024.

[64] S. Guha, A. Mullick, J. Agrawal, *et al.*, "Matscie: An automated tool for the generation of databases of methods and parameters used in the computational materials science literature," *Computational Materials Science*, vol. 192, p. 110 325, 2021.

[65] M. Chen, T. Zhang, and S. Wang, "Prompting large language models to extract chemical–disease relation precisely and comprehensively at the document level," 2024.

[66] Z. Xia, J. Ye, B. Hu, Q. Qiang, and R. Debnath, "Llmscreen: A python package for systematic review screening of scientific texts using prompt engineering," 2024.

[67] S. Liu, T. Wen, A. S. Pattamatta, and D. J. Srolovitz, "A prompt-engineered large language model, deep learning workflow for materials classification," *Materials Today*, vol. 80, pp. 240–249, 2024.

[68] H. Liu, H. Yin, Z. Luo, and X. Wang, "Integrating chemistry knowledge in large language models via prompt engineering," *Synthetic and Systems Biotechnology*, vol. 10, no. 1, pp. 23–38, 2025.

[69] Y. Zha, Y. Li, and X.-G. Lu, "Enhancing large language model comprehension of material phase diagrams through prompt engineering and benchmark datasets," *Mathematics*, vol. 12, no. 19, p. 3141, 2024.

[70] M. D. Wilkinson, "The fair guiding principle for scientific data management and stewardship: Comment," 2016.

[71] J. White, Q. Fu, S. Hays, *et al.*, "A prompt pattern catalog to enhance prompt engineering with chatgpt," *arXiv preprint arXiv:2302.11382*, 2023.

[72] H. W. Kuhn, “The hungarian method for the assignment problem,” *Naval research logistics quarterly*, vol. 2, no. 1-2, pp. 83–97, 1955.

[73] G. A. Mills-Tettey, A. Stentz, and M. B. Dias, “The dynamic hungarian algorithm for the assignment problem with changing costs,” *Robotics Institute, Pittsburgh, PA, Tech. Rep. CMU-RI-TR-07-27*, 2007.