Science and Technology of Advanced Materials: Methods

ISSN: 2766-0400 (Online) Journal homepage: www.tandfonline.com/journals/tstm20

Development of LLM-assisted data curation tools for the Starrydata materials science database

Yukari Katsura, Tomoya Mato, Yu Takada, Eiji Koyama, Dewi Yana, Atsumi Tanaka & Masaya Kumagai

To cite this article: Yukari Katsura, Tomoya Mato, Yu Takada, Eiji Koyama, Dewi Yana, Atsumi Tanaka & Masaya Kumagai (2025) Development of LLM-assisted data curation tools for the Starrydata materials science database, Science and Technology of Advanced Materials: Methods, 5:1, 2590811, DOI: 10.1080/27660400.2025.2590811

To link to this article: https://doi.org/10.1080/27660400.2025.2590811

SCIENCE AND TECHNOLOGY OF ADVANCED MATERIALS: METHODS
2025, VOL. 5, NO. 1, 2590811
https://doi.org/10.1080/27660400.2025.2590811

Taylor & Francis
Taylor & Francis Group

FOCUS ISSUE ARTICLE

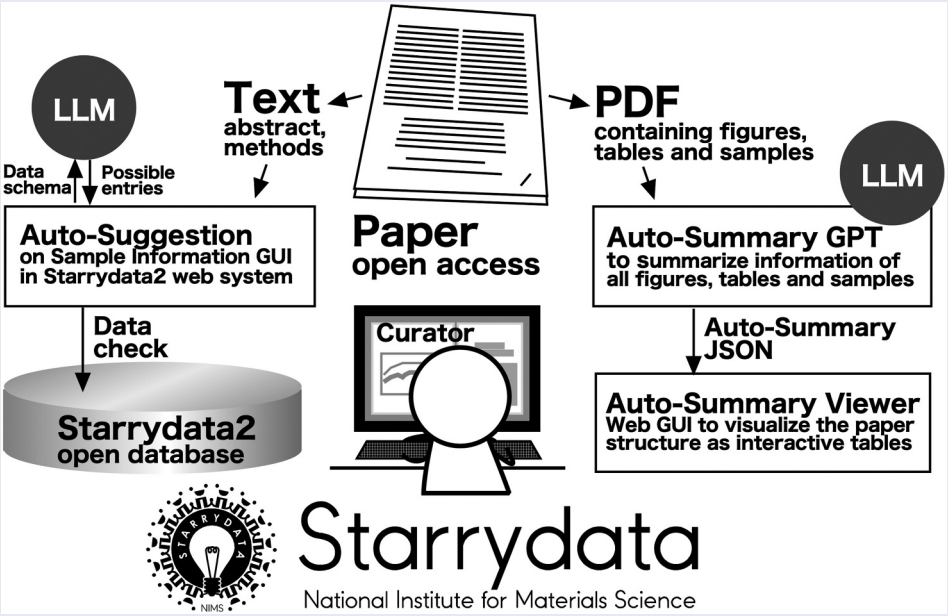
Development of LLM-assisted data curation tools for the Starrydata materials science database

Yukari Katsura[a,b,c], Tomoya Mato[a], Yu Takada[a], Eiji Koyama[a], Dewi Yana[a], Atsumi Tanaka[a] and Masaya Kumagai[d,e]

[a]Center for Basic Research on Materials, National Institute for Materials Science (NIMS), Tsukuba, Japan; [b]Graduate School of Science and Technology, University of Tsukuba, Tsukuba, Japan; [c]Molecular Informatics Team, RIKEN Center for Advanced Intelligence Project, RIKEN, Tokyo, Japan; [d]Sakura Internet Research Center, Sakura Internet Inc., Osaka, Japan; [e]Institute for Integrated Radiation and Nuclear Science, Kyoto University, Kyoto, Japan

ABSTRACT

We developed Large Language Model (LLM)-assisted tools to accelerate curator-driven data collection from materials science publications for the Starrydata database project. We implemented two complementary tools with distinct design philosophies. The first tool is the Starrydata Auto-Suggestion for Sample Information, which generates concise English descriptions conforming to our existing database schema based on user-provided text from abstracts and experimental methods. The second tool is a schema-free dual-component system comprising the Starrydata Auto-Summary GPT and Starrydata Auto-Summary Viewer. The Auto-Suggestion tool is integrated into the Starrydata2 web platform and operates efficiently with lighter language models. The Auto-Summary GPT, demonstrated using GPT-5, processes PDF files of open-access papers and generates comprehensive JSON output capturing and summarizing all figures, tables, and experimental samples as they appear in the original papers. The companion viewer transforms this JSON data into interactive tables, enabling curators to understand complete paper structure, identify relevant data collection targets, and efficiently input figure and sample information while referencing the organized data. These tools enhance curation efficiency and represent a step toward automated scientific database construction.



IMPACT STATEMENT

We developed two tools using LLMs to accelerate literature data collection in Starrydata2, including the target paper selection and the description of the figures, the samples and experimental process.

ARTICLE HISTORY
Received 9 September 2025
Revised 29 October 2025
Accepted 12 November 2025

KEYWORDS
Materials informatics; materials database; data curation; literature data mining; automated data extraction; automated knowledge extraction; large language model; data schema; web system development

CONTACT Yukari Katsura KATSURA.Yukari@nims.go.jp Materials Modelling Group, Data-Driven Materials Research Field, Center for Basic Research on Materials, National Institute for Materials Science (NIMS), 1-1 Namiki, Tsukuba, Ibaraki 305-0044, Japan

Supplemental data for this article can be accessed online at https://doi.org/10.1080/27660400.2025.2590811

1. Introduction

1.1. Complexity of materials science and the role of data-centric approaches

Materials science encompasses both inorganic and organic domains, where properties emerge from intricate interdependencies among chemical composition, synthesis conditions, crystal structure, and electronic structure [1,2]. For example, in inorganic systems, the interplay of reaction temperature and pressure can drive self-organized crystallization into complex structures, which then define the electronic states accessible to carriers. Such complexity resists simple cause–effect reasoning and often requires a combination of empirical exploration, theoretical modeling, and intuition developed through years of experience.

First-principles calculations such as density functional theory (DFT) have become essential tools, offering predictive access to electronic structures and related material properties. However, their accuracy is limited for certain key quantities. A well-known case is the band gap, where computed values can deviate substantially from experiment, directly affecting predictions of transport and optical behavior [3,4]. Consequently, while first-principles data are included in major computational repositories, their imperfect match to reality means that experimental validation and augmentation remain indispensable.

Historically, progress in experimental materials science has depended strongly on the ability of researchers to engage deeply with a chosen parent compound, iteratively adjusting chemical composition and synthesis methods to optimize properties through carrier doping, microstructure control, or defect engineering. This process reflects a broader truth: materials and their properties do not map in a one-to-one fashion. A given composition can yield diverse microstructures and thus diverse physical properties depending on how it is synthesized and processed. Experienced researchers often rely on tacit intuition that, while difficult to articulate, can outperform theory alone. Large language models (LLMs), trained on vast textual corpora and capable of reasoning flexibly over unstructured information, open a new path toward capturing and systematizing such dispersed knowledge [5,6].

By developing data-centric approaches that integrate experimental datasets with computational predictions, statistical inference, and machine learning, materials science can move beyond traditional trial-and-error discovery. The central challenge is to accelerate the process of linking synthesis, structure, and properties in a reproducible, scalable way. Materials Informatics (MI) has emerged to address this challenge, aiming to transform intuition-driven exploration into a systematic, data-driven discipline [7–12].

1.2. The Starrydata project and literature-based data collection

While computational databases have grown rapidly, large-scale experimental data remain comparatively scarce. The most widely used platforms – such as the Materials Project [13], AFLOW [14], OQMD [15], and others – have established the mainstream of materials informatics (MI) by providing crystal structures, formation energies, and electronic properties at scale, all derived from first-principles calculations. These resources remain invaluable for theory-driven exploration. However, experimental datasets, which capture synthesis variability, microstructural effects, and real-world processing conditions, have historically been more limited.

To bridge this critical gap, we launched the Starrydata project, which systematically recovers experimental datasets from published plots [16,17]. This approach enables us to share original scientific data extracted from figures without infringing on publishers' copyrights. As detailed in Ref [18], copyright grants exclusive rights to creators of original literary, scientific, and artistic works, protecting the form of the expression of ideas, but not the idea, information, or concept expressed. In the context of scientific publications, this means that while the written text and visual design of figures are protected as creative expressions, the underlying experimental data points represent factual information that cannot be copyrighted. Therefore, extracting and compiling plot data into a database is permissible, provided we do not reproduce the copyrighted forms of expression such as original text or figure images.

Unlike conventional text-mining approaches, Starrydata treats figures as primary information sources. By combining semi-automatic digitization tools with human oversight, we have curated tens of thousands of thermoelectric datapoints, ultimately creating one of the largest experimental databases for thermoelectric properties to date. Thermoelectrics were chosen as a proof of concept because of the field's rich scientific interdependencies – between Seebeck coefficient, electrical resistivity, and thermal conductivity – and its popularity within MI, where these complexities make it an attractive testbed for data-driven discovery.

Building on this foundation, Starrydata has expanded beyond thermoelectrics. Collaborations have enabled the construction of a large-scale dielectric property dataset (in partnership with Murata Manufacturing Co., Ltd.) [19], systematic datasets on quasicrystals and their approximants through the HyperMaterials project [20], and ongoing work on magnetic materials [1]. These demonstrate that the Starrydata framework is broadly generalizable across experimental domains.

Several related efforts highlight the broader landscape of literature-based data collection. Classical curation projects, often subscription-based, built reliable datasets through meticulous manual work but limited their accessibility. Community-driven efforts for large-scale open data curation are reported for thermophysical properties [21], polymer nanocomposites [22], and perovskite solar cells [23] Open experimental repositories built via high-throughput thin-film experiment is reported [24]. These diverse projects illustrate the many possible routes to experimental data infrastructures in MI.

Starrydata's contribution lies in its distinctive focus on figure-derived data. Semi-automatic tools such as WebPlotDigitizer [25] and our StarryDigitizer [1] have proven indispensable, allowing curators to interactively refine extraction algorithms, while fully automated approaches – ranging from classical image analysis to transformer-based systems such as DePlot [26] and MatGD [27] –remain promising but are not yet mature for practical deployment. Importantly, our experience indicates that figure digitization, while non-trivial, is seldom the true bottleneck in database construction. The more labor-intensive step is metadata labeling: for each digitized dataset, one must record, in a unit-convertible format, the sample composition, fabrication method, and measurement conditions, to ensure that the extracted numbers are meaningful for downstream analysis.

In this context, existing text-mining pipelines such as ChemDataExtractor [28] and related text-mining projects [29–37] have demonstrated how large volumes of literature can be converted into structured databases. Their strength lies in capturing explicitly described entities and relations at scale. However, because they rely on predefined vocabularies, they may miss key information when terminology differs or when relevant details are only implicit, leading to gaps in datasets. LLMs offer a promising complement: by reasoning over broader textual context, they can infer missing or implicit information – such as synthesis routes or measurement definitions – that are not always stated in canonical form. This capability opens a path toward extracting richer, more complete metadata, thereby improving both the coverage and usefulness of experimental databases.

1.3. LLMs and metadata representation in materials informatics

Recent years have witnessed a surge of interest in applying large language models (LLMs) to materials informatics. Reported applications span diverse directions, including benchmarking knowledge of materials science concepts, developing agentic systems for autonomous discovery, and supporting human–machine collaboration [38–43].

Of particular relevance to this work are projects that use LLMs for literature-based data extraction. Tools such as GPTArticleExtractor [44], MaTableGPT [45], MatterChat [46], and domain-specific implementations for thermoelectric materials [47], magnetic materials [48] and superconductors [49] have demonstrated that structured datasets can be obtained directly from research papers. These systems differ in scope and technical design – for example, some employ multi-step prompting strategies that iteratively query documents, while others aim for one-shot extraction tailored to predefined schemas. Collectively, they illustrate the feasibility of integrating LLMs into database-building workflows, while also highlighting the diversity of approaches still under active exploration.

In parallel, the representation of extracted knowledge has been a longstanding concern in materials informatics. Ontology-based frameworks – spanning early proposals for experiment metadata, catalyst databases, and more recent schema development – seek to provide consistent vocabularies and hierarchical structures for recording synthesis, processing, and measurement details [50–53]. Although independent of LLMs, these frameworks address the same bottleneck: ensuring that experimental data are captured in formats amenable to integration, reuse, and reasoning. When embarking on data collection in a new field, existing ontologies can guide the design of what metadata items should be recorded. Then LLMs can be used to suggest relevant entries for these predefined metadata categories, streamlining the most labour-intensive step of database construction.

1.4. Limitations and legal considerations

Despite these advances, fully automating data collection in materials science remains infeasible. Numerical values extracted from graphs must be contextualized with metadata – composition, processing method, measurement details – recorded in unit-convertible formats. This step, rather than digitization itself, constitutes the major bottleneck. While LLMs offer promising assistance, they cannot yet replace the flexible judgment of human curators. Moreover, legal restrictions sharply limit their deployment. As of 2025, major publishers such as ACS [54], Elsevier [55], and Wiley [56] explicitly prohibit applying AI tools to their journal content under e-journal terms of use. Even when text and data mining (TDM) licenses can be purchased, transmitting such licensed content to high-performance commercial LLMs hosted online is often contractually restricted and raises data security concerns. These realities make universal LLM-based automation impossible in the current publishing environment.

The Starrydata project therefore adopts a hybrid approach. Human data collectors – valued for their meticulous work – remain central, while automation is introduced where legally and technically permissible. Papers from publishers allowing AI use can serve as training exemplars, improving both LLM-assisted tools and the skills of human curators. Until publishing policies evolve, this hybrid strategy provides a legally sound and practically effective path forward.

Currently, the curators and the users of the Starrydata upload the experimental data in the plots, typically in the following procedure. Firstly, they find a paper that contains the target data. When they send the DOI of the paper, Starrydata2 web system retrieves the bibliographic information from CrossRef and add the paper to our database. Secondly, they select a target figure in the paper. They input the information of a plot in a figure, by inputting the figure name (number), selecting the physical properties and the units for the horizontal (x) and the vertical (y) axes. Thirdly, they define a sample in the plot. Such samples can appear in multiple plots in the paper. Sample name is usually given from the label in the plot, but it is usually not informative enough. Therefore,

In this study, we explore how LLMs can accelerate metadata extraction within the Starrydata workflow. Using thermoelectric materials as an example, we evaluate existing text-mining and LLM-based methods, compare prompts, models, and schema designs, and integrate an LLM-assisted function into Starrydata2. Although applicable only to open-access publications, this represents a significant step toward more efficient and scalable experimental data collection.

2. Methods

We developed two complementary LLM-assisted tools for data curation. The first tool, Starrydata Auto-Suggestion for Sample Information, was implemented as an integrated feature within the existing Starrydata2 web system [1], utilizing users' registered OpenAI API keys to generate schema-compliant sample descriptions. The second tool consists of two components: the Starrydata Auto-Summary GPT, developed using OpenAI's custom GPT platform and distributed through the GPT Store, and the Starrydata Auto-Summary Viewer, created as a standalone web application that processes JSON output from the GPT and presents it through interactive tables. These two tools serve distinct purposes, with the Auto-Suggestion tool integrating directly with the database interface for streamlined data entry, while the Auto-Summary tool system operates independently to extract and organize comprehensive information from scientific publications. Detailed implementation and technical specifications for each tool are provided in the following sections.

3. Starrydata auto-suggestion for sample information

In Starrydata's literature data collection [1,17], a curator or user first specifies a paper for data collection. When a DOI is provided, the Starrydata2 web system automatically retrieves bibliographic information from CrossRef [57] and generates a set of links and pages for data collection. A screenshot of a plot can be taken from the PDF accessed through the 'Original paper' link, which directs to the publisher's website specified by the DOI. After pasting the screenshot into the digitization interface (WebPlotDigitizer [25] or our original StarryDigitizer) embedded in the 'Data' page, numerical data for each curve can be extracted using semi-automatic algorithms. This numerical data is then entered into the textbox at the top of the same page. To save each curve in the database, curators must specify the figure name, summarized caption, and physical quantities for both x and y axes, including their exponents and units. Additionally, they must specify the sample, including its molar composition. Each paper typically contains multiple samples, and a single sample may appear across different figures. More detailed sample information – such as synthesis methods, forms, grain sizes, and measurement methods – can be added through project-specific input fields on a separate 'Sample information' page.

While graph tracing is often seen as the main task in this process, the most time-consuming aspect for curators is describing the samples. This requires reading the main text, identifying and understanding each sample, and filling in detailed information for every sample. Without proper sample identification, even extensive numerical data has limited value – we need to know which sample and measurement method produced each dataset. Furthermore, there is a risk of accidentally including non-experimental data, such as theoretical calculations, if samples are not properly characterized.

Database projects tend to define input items for all 'information that might be needed in the future' and 'all factors that might influence properties'. However, increasing the number of items directly increases the burden on data collectors. Entering explicitly stated information is relatively quick, but for information not explicitly stated, data collectors tend to search exhaustively until they find it. If items with a low likelihood of being stated are made mandatory, data collection time increases substantially and long-term motivation declines. To sustain multi-year data collection, data design that maintains a high success rate is essential.

To address this, Starrydata have defined a minimal, project-specific data structure for each collection

project. Figure 1 shows the initial data structure used in the thermoelectric-property data collection project, showing the keys and the lists of categories for the sample information. Each key contains two fields: 'category' and 'comment' in the JSON. The category field is a preconfigured set of choices based on domain expertise, and the data collector selects the option judged most appropriate from a dropdown. This approach completes normalization during data entry and yields datasets suitable for machine learning. However, judging the appropriate choice requires substantial expertise. In data collection projects staffed primarily by non-specialists, some efforts have been abandoned due to the difficulty.

In this work, we address semantic ambiguity by providing the target data schema directly to the LLM. This ensures that metadata keys are consistent and clear. While classical text mining programs have difficulty with merging related concepts, LLMs are very good at understanding context and finding relationships between different terms – often better than human curators.

3.1. The original sample information schema in Starrydata

Figure 1 shows the current data schema used from the beginning of the Thermoelectric Materials project in Starrydata2 web system. Designed in 2018, this data structure admits various variants depending on the research domain and interests. To avoid time-consuming searches for the desired item when categories are too numerous, we have limited the number of the categories. We also needed to design categories mutually exclusive as much as possible, to reduce time spent deliberating among multiple choices.

These constraints lead to suboptimal aspects in the data structure. For example, the 'Pressure sintering' category treats together both current-assisted sintering methods (variously called spark plasma sintering (SPS), pulsed current sintering, etc., depending on manufacturer) and hot pressing without current. Practically, these methods deliver dense polycrystalline bulks under mild pressure in a graphite die or similar, and the resulting bulk properties are similar. Nevertheless, for data collectors without such experimental experience, accurate classification based solely on device or method names in the paper is difficult.

Moreover, many synthesis processes are sequences of multiple techniques, making a single-choice selection difficult. We instruct collectors to prioritize the final process. For example, when an electric furnace anneal follows pressure sintering, 'Pressure sintering', which more strongly governs properties, should be selected rather than 'annealing'. Such choices retain elements that depend on empirical judgment.

StructureType
Bi2Te3
PbTe
Skutterudite
Clathrate
d-silicide
Silicide
Si
Perovskite
Chalcogenide
SnSe
Cobaltite
Antimonide
Organic
ZnO
Heavy-Fermion
Half-Heusler
Heusler
Alloy
Sulfide
Oxide
Zintl
Selenide
Phosphide
Boron
Telluride

FabricationProcess
MeltGrowth
FluxGrowth
VaporGrowth
FilmGrowth
Melting
MeltQuench
HighPressureSynthesis
PressureSintering
Sintering
MeltQuench
MechanicalAlloying
SoftChemical
Hydrothermal
Organic

Form
SingleCrystal
Bulk
OrientedBulk
EpitaxialFilm
Film
Ribbon
Wire
Device
Module

Purity
SinglePhase
Impurities
Precipitates
MetalComposite
Composite

ElectricalMeasurement
SteadyState+4P
SteadyState+2P
SteadyState+vdP
Harman
MicroProbe

ThermalMeaasurement
LaserFlash+DSC
LaserFlash
LaserFlash+3R
LampFlash+DSC
LampFlash
LampFlash+3R
Harman
Disc
3omega
ThermoReflectance

GrainSize
>= 1 cm
>= 1 mm
>= 100 micron
>= 10 micron
>= 1 micron
>= 100 nm
>= 10 nm
< 10 nm

RelativeDensity
>=95%
>=90%
>=80%
>=70%
>=60%
>=50%
>=40%
>=30%
<30%

Figure 1. Initial data schema for sample information entry used in the thermoelectric materials project of the Starrydata2 web system.

Evaluating thermal conductivity requires values for both thermal diffusivity and specific heat. Even when thermal diffusivity is measured by the laser flash method, methods for obtaining specific heat vary by researcher, including DSC measurements, measurements using a reference sample within the laser flash instrument, three times the gas constant R based on the Dulong–Petit law, or literature values. These methods are often not explicitly stated in the experimental section. While it would be more elegant to separate 'thermal diffusivity measurement method' and 'specific heat measurement method' into two categories, doing so would substantially increase input time. We therefore define categories for commonly used combinations, though not all combinations are covered.

There are also challenges from dynamic category changes. Since launch, new categories have been continually added, and user-added categories also appear in the dropdown. Because user additions often reproduce the names used in the paper, some overlap with existing categories, compromising the initial exclusivity. Budget constraints have prevented us from establishing a process and staffing to rigorously curate new categories in close discussion with users.

To capture details that cannot be expressed in the category field, each item includes a free text 'comments' field. Data collectors can record succinct summaries of details in English, but this demands subject-matter knowledge and concise English expression. It is difficult to recruit personnel who possess these skills and can tolerate the monotony of data collection.

To address these issues, we developed an LLM-based function that proposes entries for the comments field. LLMs possess vast domain knowledge and can compensate for collectors' knowledge gaps. They excel at summarization tasks, strongly supporting the collector's reasoning process. Paper authors often employ creative names to emphasize novelty; even when a method is nearly identical to an existing one, collectors may hesitate in normalization. In such situations, an LLM's ability to make a holistic judgment and normalize to a more general expression is highly beneficial for database builders. Moreover, in projects primarily staffed by non-native English speakers, an LLM that produces natural English represents significant support.

3.2. Development of the LLM-based auto-suggestion system

A screenshot of the Sample Information GUI in Starrydata2 web system is shown in Figure 2. The Sample Information GUI augments the long-used interface with an 'Auto-extraction' toggle. When turned ON, a text box appears at the top of the Sample Information GUI. Users register their OpenAI API key in the profile management page and paste the paper's abstract and experimental methods paragraphs as plain text. Through the API, the text and the list of target descriptors are sent to an external LLM, and JSON-formatted parsing results appear in green text beneath each input field. After reviewing the output, the data collector can click to auto-fill fields judged correct. To aid reviewing the automatic suggestion, a text box on the right highlights the words in the source text that matches the words in the suggestions by the LLM.

By default, the system attempts automatic retrieval for all sample descriptors, but one can specify descriptor IDs to improve efficiency. The LLM function is limited to open-access papers or papers from publishers that allow sending content to LLMs, and a cautionary note is displayed in the GUI.

When using this auto-suggestion feature, the curator simply pastes paragraphs from experimental methods (and the abstract), presses the button, and reviews the suggested entries for each field. They can accept correct suggestions with a click, fill in any remaining fields, and save all information. This can dramatically speed up the sample information description process, though the actual improvement depends on the curator's approach and confidence in the LLM. For example, one trained collector chose to re-read the entire paper and verify all LLM suggestions before registration, resulting in minimal throughput improvement. However, even in such cases, having the LLM as an expert 'consultant' reduces the burden of writing concise English summaries and eases the overall workload. The support effect is expected to be even greater for collectors who previously avoided sample entry due to limited expertise, or those who are more willing to trust and use LLM outputs effectively.

4. Starrydata auto-summary GPT/viewer

The sample-information collection based on our custom data schema in the previous section was designed to reduce the burden on data collectors; consequently, it captures only a small subset of the information present in a paper. We therefore developed a tool – premised on the use of LLMs – that collects more detailed information in a concise, database-ready form.

For every figure in a paper, the tool automatically retrieves a caption summary, the physical quantities and units on each axis, and supplementary information to aid understanding. For every sample appearing in the paper, it automatically retrieves chemical composition, crystal structure of the parent compound, a synthesis summary, and concise insights from phases and microstructures. To represent this information, we designed a unified, general-purpose JSON schema. We then created a custom GPT on

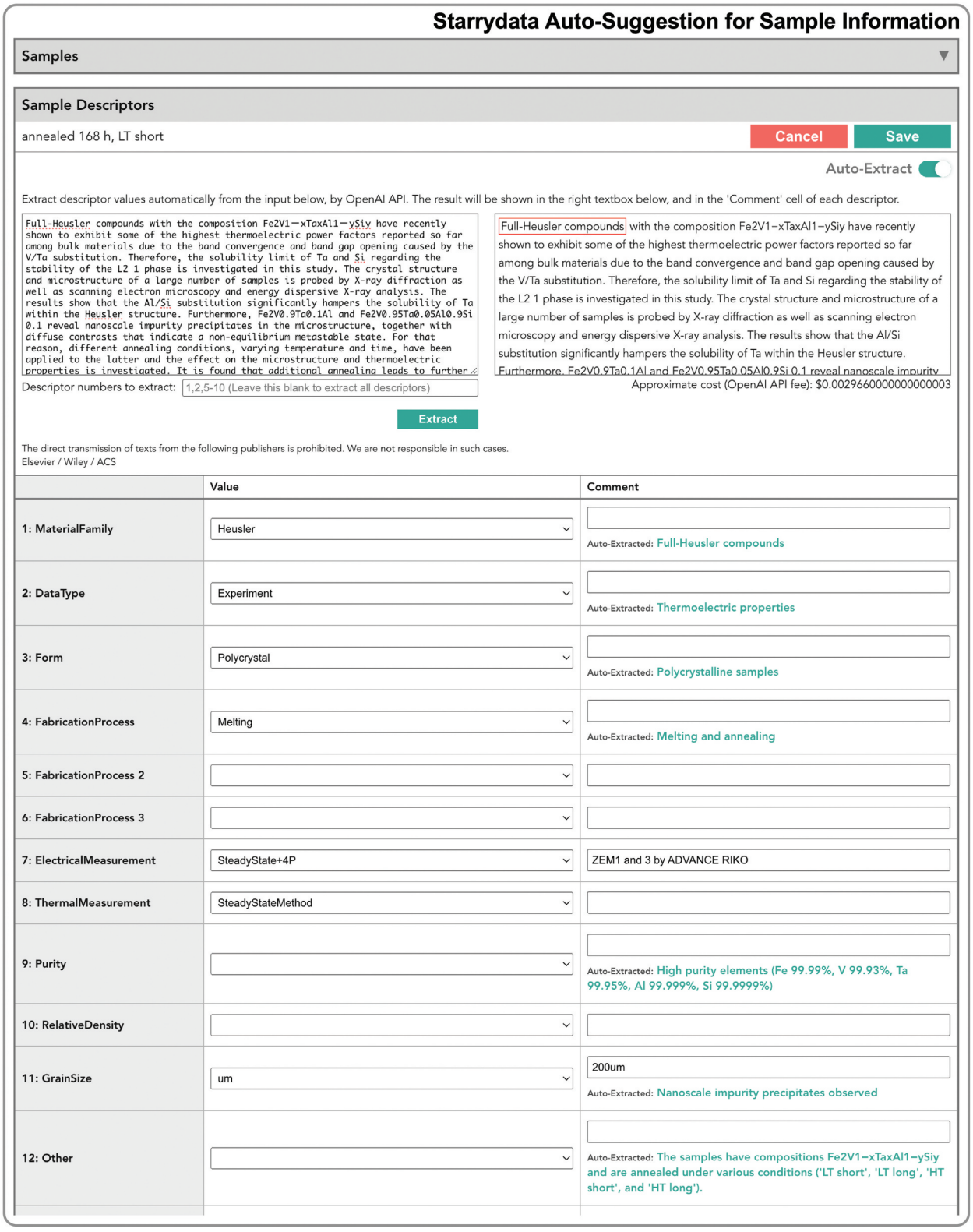


Figure 2. Starrydata auto-suggestion for sample information feature integrated into the sample information GUI of the Starrydata2 web system. The screenshot shows the interface after pasting abstract and experimental method text extracted from reference and clicking the suggestion button.

OpenAI that, upon paper upload, outputs JSON conforming to the schema. We also prepared a tool that visualizes the resulting JSON as readable tables in a web browser.

At present, we do not provide a mechanism for data collectors or users to register these JSON outputs into the Starrydata database for access by others. This is because the creation of the JSON effectively requires the use of an LLM, while only a small fraction of papers permit LLM use. We cannot adjudicate whether users have inadvertently uploaded papers to an external LLM in violation of terms of use, and we

therefore avoid the risk of publishing such data as part of our database.

4.1. Design of a general JSON schema

To systematically organize bibliographic information on experimental thermoelectric materials, we designed a unified JSON schema. As shown in the Supplementary Material, the schema is divided into four main areas.

The JSON schema was designed in a flexible form, which will be applicable for various papers in materials science. The keys of the first layer are 'paper_metadata', 'scope_check', 'figures', 'tables', 'samples_common', 'samples', and 'provenance. The 'scope_check' key contains the short text for 'domain', 'purpose' and 'approach' summarized by the LLM, and the Booleans named 'is_original_research', 'is_review', 'is_experimental' and 'is_theoretical'. The 'figures' key contains a set of subfigures with summarized caption. For graphs, the information of x and y axes, including the physical quantity, unit and its conversion equation to SI unit system, reference ticks and scale type (linear or log). The extraction of data points did not work well with currently available LLM models, so did not include the key for extracted data in the JSON schema. The 'tables' key was designed to describe the tables, including the headers and rows. The information about experimental methods and results were summarized in 'samples_common' and 'samples' keys. The structures of these two keys are similar, and 'samples_common' key contains the common information all over the samples, and 'samples' key contain overriding sample-specific information. The compositions (starting/target/analyzed) keys are only given under the 'samples' key. The 'provenance' key contains the information about this curation, such as extraction method, annotator and date. These fields will be filled by the Starrydata web system in future.

This JSON schema integrates paper information and experimental data to ensure machine readability and interoperability. In thermoelectric materials research, it is especially useful for facilitating cross-study comparisons and meta-analyses.

4.2. Starrydata auto-summary GPT

The 'Starrydata Auto-Summary GPT' is a specialized tool built using OpenAI's custom GPT feature, which was introduced in late 2023. Custom GPTs allow users to create tailored versions of ChatGPT by providing specific instructions, knowledge, and behavioural guidelines that persist across conversations. Unlike standard ChatGPT interactions where users must repeatedly provide context and instructions, a custom GPT retains pre-configured prompts and domain-specific knowledge, ensuring consistent behaviors for specialized tasks. In our implementation, we configured the custom GPT with detailed instructions for extracting and formatting sample information from scientific papers, including specific data schemas and output formats required by the JSON above. This approach enables any curator to access a consistently configured AI assistant without needing to understand or manage the underlying prompts themselves.

The prompt used for Starrydata Auto-Summary GPT is shared in the Supplementary material, with an example JSON output. We optimized the prompt by specifying the information to be extracted and adding cautions on common pitfalls, and we provided the general JSON schema to elicit correctly formatted JSON. Because visualization fails if the JSON is malformed, we also prepared a prompt for self-checking the format.

The prompt for Starrydata Auto-Summary GPT instructs the LLM how to respond to the user input by using natural language. This prompt consists of 8 sections: (1) Introduction, (2) Inputs, (3) Task, (4) Output requirements, (5) JSON schema, (6) Process guidelines, (7) Ambiguity and missing data, and (8) Final deriverables.

We firstly defined the character for the LLM by stating *You are a meticulous research assistant specialized in extracting structured information from experimental materials science papers. Your job is to read a provided open-access paper and produce a strictly formatted JSON according to the schema and rules below*. Then we defined the acceptable forms of Inputs as: '*The user attaches a PDF file of a paper. Before processing any attached PDF, you must verify that the paper is open access, a preprint such as arXiv, or an author's accepted manuscript that has been made publicly available. Never process subscription-only or paywalled publisher PDFs*', to avoid the violation of the Publishers' terms of use. In the Task section, we explained what to do as follows. *'From the paper, determine whether it is an experimental materials science study where original samples were synthesized, processed, and characterized. Perform a scope check to identify the research domain, whether the paper is original research or a review, whether it contains experimental work, theoretical work, or both, and to capture the paper's purpose and approach in concise free text. When both experimental and theoretical components are present, both flags must be true. Extract figures and subfigures including graph axis details. Extract all tables including headers and rows. Extract all samples and their compositions including parent, starting, target, and analyzed compositions with molar ratios. Extract microstructure and methods. Produce a JSON object exactly following the schema and constraints'*. The

details of the output forms were separated in the Output requirements section.

After giving the JSON schema, we gave the Process guidelines section how to process the given paper and what information to focus on. Apart from these guidelines, the most important messages were emphasized by giving as separate sections. The Ambiguity and missing data section is a short section to avoid hallucination, by stating '*If a field cannot be determined, leave it null or empty and include a short explanatory note in the notes or other fields. Do not invent compositions or values under any circumstances.*' The Final Deliverables section is also a short section stating, '*Return data.json containing only the extracted JSON, formatted inside a code block labeled data.json.*'

To prevent users from inadvertently uploading papers to an external LLM in violation of e-journal terms, the workflow first requests the DOI. Only if no policy issues are detected does it instruct users to upload the file. The prompt will undergo minor iterative refinements during operation to improve both output accuracy and compliance.

4.3. Starrydata auto-summary viewer

We developed a web GUI to present JSON produced by the Starrydata Figure/Sample Describer in an easy-to-read format. Currently, it is a single-page web application independent of the Starrydata2 system, available at https://auto-summary.starrydata.org/. Researchers simply paste the JSON into a text area and click the 'Render Tables' button to visualize it immediately.

The display includes bibliographic metadata; a scope check determining whether the paper is an experimental study that measures samples prepared by the authors (as required for our data target); the research domain; caption summaries and axis information for each figure and subfigure; properties and measurement conditions common to all samples; individual sample information and overrides; and provenance of data extraction. The tables are compactly arranged, with interactive expansion for details as needed. The viewer also offers the ability to load minimal sample data and to save the current view as an HTML file, facilitating data preservation and reproducibility. The tool is positioned as an aid for visualizing quantitatively organized information from papers to support efficient comparison and analysis by researchers.

4.4. Example: analysis of a Heusler thermoelectrics paper

As a use case, we submitted our open-access paper on Heusler thermoelectrics to the custom GPT (Starrydata Auto-Summary GPT) and visualized the extracted information about all figures, all tables, and all physical samples paper. The results are shown in Figures 3, 4, respectively. A paper on Heusler thermoelectric materials, containing both the theory and the experiment, is used as an example.

As indicated at the top of Figure 3, in addition to figures commonly found in such papers – powder XRD patterns, microstructural images, and temperature-dependent thermoelectric property graphs – the paper includes scatter plots comparing multiple samples, first-principles (theoretical) calculations, and schematic diagrams explaining physical phenomena, covering a diverse range of figure types typical in thermoelectric papers. The analysis used ChatGPT-5, released in August 2025. This model is reported to outperform GPT-4-series models in image recognition and to deliver accurate judgments via chain-of-thought reasoning.

The time from submitting the PDF to the Starrydata Auto-Summary GPT until the target JSON was returned likely varies with ChatGPT server load and tuning. With ChatGPT-5, it was 3 minutes 48 seconds (37 seconds for thinking, 3 minutes and 11 minutes to output JSON).

The lower part of Figure 3 shows the extracted Paper Metadata, Scope Check, and Figures sections. Paper Metadata was retrieved accurately, and the Notes field included license information (Open Access, CC BY 4.0) and a concise summary of the study that is not evident from the title alone. The Scope Check indicated that the paper is within scope for the data collection project, enabling data collectors to decide immediately whether to proceed with data collection.

The Figures section listed 'Figures 1–7' and provided subfigure information where applicable. Each caption was concisely summarized, and insights obtained from each figure were recorded based on the paper's claims. The 'Graphs' column indicated the number of graphs in each figure – either 0 or 1 in this paper. Expanding the collapsed 'Graphs-Details' revealed the physical quantities and units on the × and y axes, representative tick values, linear/log scaling, and conversion formulas to SI units. These were inferred from figure captions, main text, and image recognition, forming a highly useful table for data collectors to grasp a paper's overall structure and plan collection.

Figure 4 shows the Samples section, where sample information appearing in the paper is presented. As instructed by the prompt, the information is divided into common and per-sample entries, yielding tables with minimal redundancy and clear sample-by-sample differences. Whereas data collectors had added three samples by reading the paper's graphs, the analysis identified four additional entries – for a total of seven.

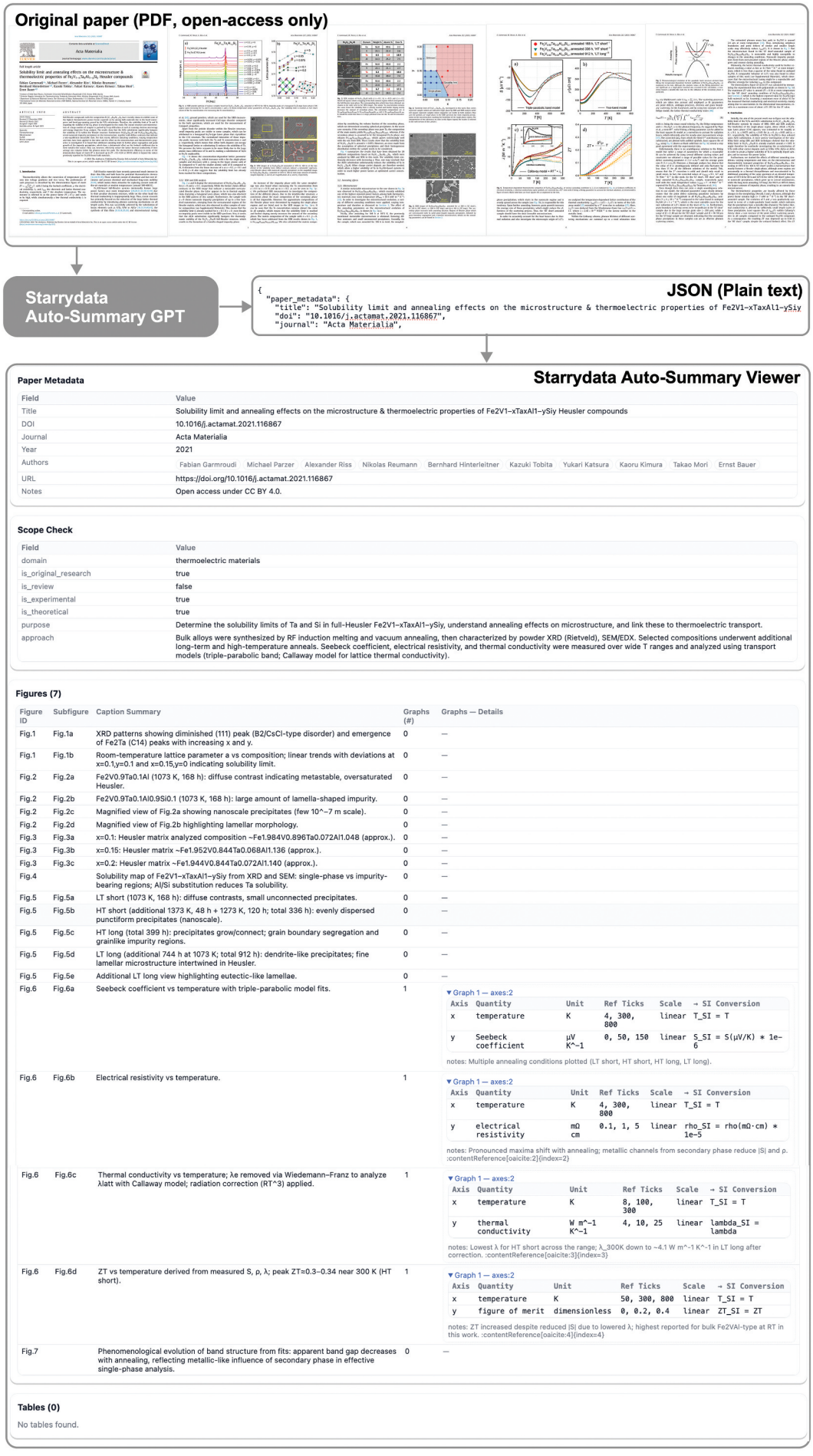


Figure 3. Workflow demonstration of Starrydata Auto-summary GPT/Viewer. Screenshot of the display generated by pasting the JSON output from Starrydata Auto-summary GPT (obtained by submitting the PDF of the original paper [59]) into the Starrydata Auto-summary Viewer. Due to space limitations, only the metadata, scope, figures, and tables sections are shown.

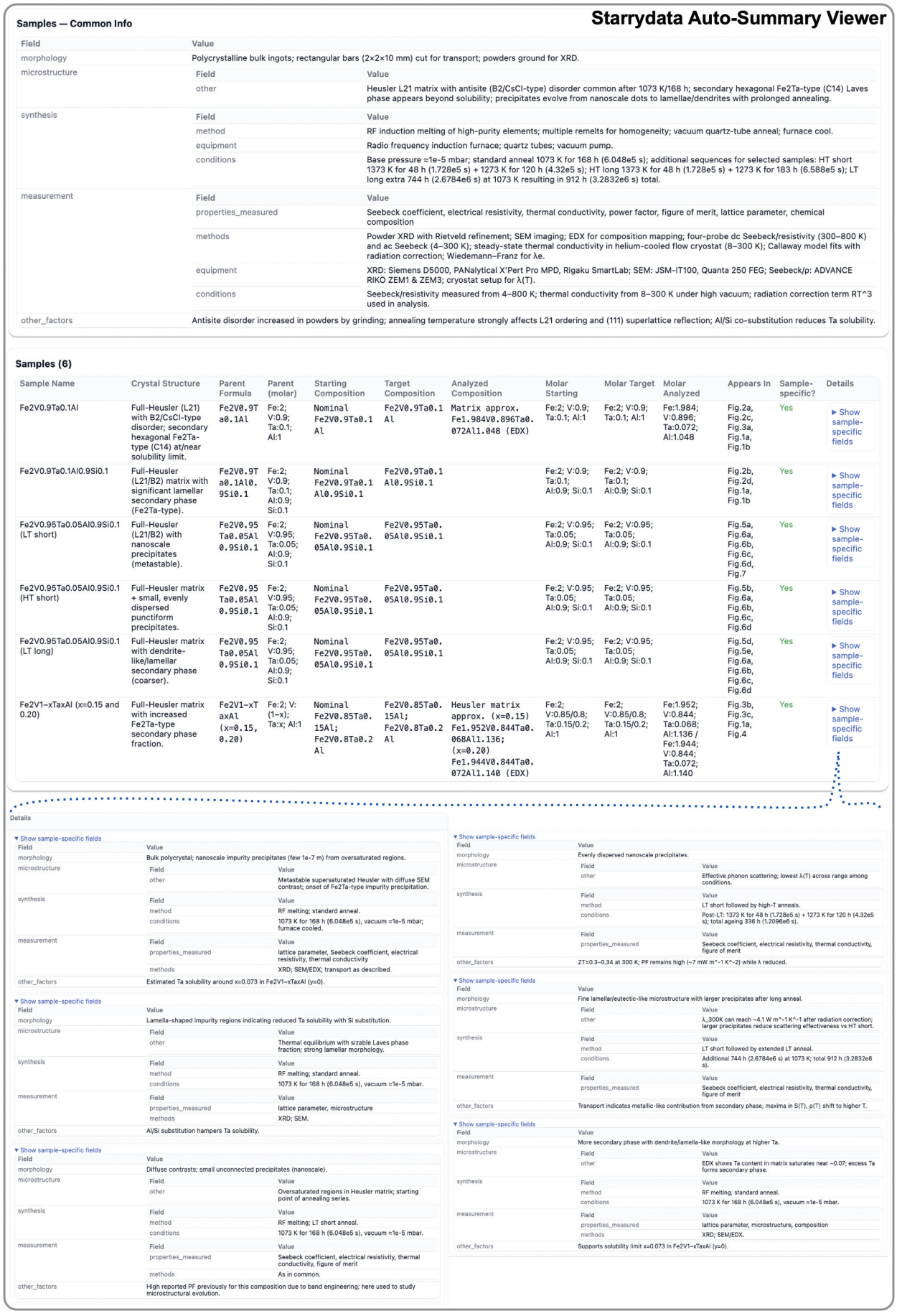


Figure 4. Screenshot of the samples table that could not be fully captured in Figure 3. The table displays common information and compositional data for each sample. The detailed sample-specific information, which appears when clicking 'show sample-specific fields' within the samples table, is shown separately below due to space constraints in the figure layout.

These four additional entries correspond to phases or regions confirmed by XRD and microstructural observations: one pure target $L2_1$/B2 phase and three microstructures containing a secondary Fe_2Ta (C14) phase with different microstructures. The last three could be merged with the first three based on composition and may be considered the same sample in context. If data collectors wish to reflect such improvements, they can provide feedback to the Starrydata Figure/Sample Describer and regenerate the JSON; such interactive iteration can lead to better paper descriptions.

Given the variety of definitions of chemical composition, we instructed the system to output starting, target, and analyzed compositions. A 'Show sample-specific information' section to the right of each sample reveals fields such as morphology, microstructure, synthesis, and measurement when expanded. These are free-form texts summarizing synthesis methods concisely at a granularity that allows materials scientists to envision concrete procedures. By not enforcing an overly rigid data structure, we gain high information density. Listing these texts and resubmitting them to an LLM can produce structured tabular data tailored to user interests, making the information easier to use in various formats, including machine learning.

Independent validation using Claude Opus 4.1 [58], another state-of-the-art LLM like ChatGPT, revealed that the JSON extraction achieved 92.2% precision (331 out of 359 extracted key-value pairs were completely correct) with an additional 5.6% being partially correct, yielding 97.8% overall precision for the extracted data. The systematic errors were exclusively related to image interpretation: micro-prefixes (μ) in unit notations from figure axes were not captured, resulting in simplified units ('V K^{-1}' instead of 'μV K^{-1}', 'Ω m' instead of 'μΩ cm'). Additionally, the reference_ticks field was populated with significant measurement values from the paper text (e.g. thermal conductivity values of 4.1, 6.7, 7.1, 9.1, 25 W m^{-1} K^{-1}) rather than actual axis tick marks, indicating that when axis ticks could not be visually recognized, the model substituted contextually relevant numerical data from the text. No errors were found in text extraction, chemical formulas, or experimental conditions. While recall cannot be quantitatively assessed since it depends on the intended depth of extraction, any additional information not initially included can be readily obtained by requesting ChatGPT to update the JSON through continued conversation.

5. Conclusion

To enable efficient data collection from materials science papers, we developed two LLM-based assistance tools designed for open-access publications. The first tool is Starrydata Auto-Suggestion for Sample Information, implemented within the Starrydata2 web system, which automatically generates descriptive comments for sample information keys based on project-specific data schemas from plain text extracted from paper abstracts or experimental sections by users. The second tool is Starrydata Auto-Summary GPT/Viewer, which comprehensively summarizes figures, tables, and sample information from user-provided PDFs, outputting structured data in a unified JSON schema for visualization in the Viewer interface.

These tools are designed to assist and streamline the work of human data collectors rather than replace them entirely. Given current constraints imposed by publisher terms of use, human understanding and summarization of scientific papers remain essential. Should these constraints be relaxed in the future, the technologies developed here could be applied more broadly and are expected to significantly accelerate data-driven research in materials science.

Acknowledgements

We express our gratitude to our data curators, anonymous crowd workers and the voluntary users of Starrydata for sharing their extracted data. We deeply thank all the collaborators from academia and industries for financial supporting our staffs including the data curators. We are grateful to all those who provided valuable advice and suggestions throughout this work, and to those who offered opportunities for disseminating our research.

Author contributions

Y. Katsura developed the Starrydata Auto-Summary GPT/Viewer and wrote the main manuscript. Starrydata2 web system and the Auto-Suggestion function was developed by T. Mato, Y. Takada and M. Kumagai. Application of LLMs for data collection was explored by T. Mato, E. Koyama and D. Yana. The Starrydata project was managed by A. Tanaka and Y. Katsura.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This research was supported by JST-CREST [Grant JPMJCR19J1], the Kazuchika Okura Memorial Foundation, and our industry partners.

Data availability statement

The Starrydata2 web system including the Starrydata Auto-Suggestion for Sample Information can be accessed from https://www.starrydata2.org. The Starrydata Auto-Summary GPT is distributed in the OpenAI GPT Store at https://chatgpt.com/g/g-68a6737fa61881919561

c553e5aac277-starrydata-auto-summary-gpt. The prompt used in the GPT is shared in the supplementary material. The Starrydata Auto-Summary Viewer can be accessed from https://auto-summary.starrydata.org, and the source code is distributed at https://github.com/starrydata/auto-summary.

References

[1] Ashcroft NW, Mermin N. Solid state physics. Florence (KY): Brooks/Cole; 1976.

[2] Borlido P, Schmidt J, Huran AW, et al. Exchange-correlation functionals for band gaps of solids: benchmark, reparametrization and machine learning. npj Comput Mater [Internet]. 2020;6(1). doi: 10.1038/s41524-020-00360-0

[3] Katsura Y, Takagi H, Kimura K. Roles of carrier doping, band gap, and electron relaxation time in the Boltzmann transport calculations of a semiconductor's thermoelectric properties. Mater Trans. 2018;59(7):1013–1021. doi: 10.2320/matertrans.E-M2018813

[4] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Adv Neural Inf Process Syst [Internet]. 2017;30:5998–6008.

[5] Naveed H, Khan AU, Qiu S, et al. A comprehensive overview of large language models [Internet]. ArXiv [cs.cl]. 2023. Available from: http://arxiv.org/abs/2307.06435

[6] Jain A, Hautier G, Ong SP, et al. New opportunities for materials informatics: resources and data mining techniques for uncovering hidden relationships. J Mater Res. 2016;31(8):977–994. doi: 10.1557/jmr.2016.80

[7] Butler KT, Davies DW, Cartwright H, et al. Machine learning for molecular and materials science. Nature. 2018;559(7715):547–555. doi: 10.1038/s41586-018-0337-2

[8] Imran QF, Kim D-H, Bong SJ, et al. A survey of datasets, preprocessing, modeling mechanisms, and simulation tools based on AI for material analysis and discovery. Materials [Internet]. 2022;15(4):1428. doi: 10.3390/ma15041428

[9] Choudhary K, DeCost B, Chen C, et al. Recent advances and applications of deep learning methods in materials science. npj Comput Mater. 2022;8(1):1–26. doi: 10.1038/s41524-022-00734-6

[10] Wang Z, Chen A, Tao K, et al. MatGPT: a vane of materials informatics from past, present, to future. Adv Mater. 2023;36(6):e2306733. doi: 10.1002/adma.202306733

[11] Van M-H, Verma P, Zhao C, et al. A survey of AI for materials science: foundation models, LLM agents, datasets, and tools [Internet]. arXiv [cs.LG]. 2025. Available from: http://arxiv.org/abs/2506.20743

[12] Jain A, Ong SP, Hautier G, et al. Commentary: the materials project: a materials genome approach to accelerating materials innovation. APL Mater [Internet]. 2013;1(1). doi: 10.1063/1.4812323

[13] Curtarolo S, Setyawan W, Wang S, et al. Aflowlib.org: a distributed materials properties repository from high-throughput ab initio calculations. Comput Mater Sci. 2012;58:227–235. doi: 10.1016/j.commatsci.2012.02.002

[14] Saal JE, Kirklin S, Aykol M, et al. Materials design and discovery with high-throughput density functional theory: the open quantum materials database (OQMD). JOM. 2013;65(11):1501–1509. doi: 10.1007/s11837-013-0755-4

[15] Katsura Y, Kumagai M, Mato T, et al. Starrydata: from published plots to shared materials data. Sci Technol Adv Mater Methods [Internet]. 2025;5(1). doi: 10.1080/27660400.2025.2506976

[16] Cotton FA, Wilkinson G, Mutillo C, et al. Advanced inorganic chemistry, 6th edition. J Chem Educ. 2000;77(3).

[17] Katsura Y, Kumagai M, Kodani T, et al. Data-driven analysis of electron relaxation times in PbTe-type thermoelectric materials. Sci Technol Adv Mater. 2019;20(1):511–520. doi: 10.1080/14686996.2019.1603885

[18] Elliott R. Who owns scientific data? The impact of intellectual property rights on the scientific publication chain. Learn Publ. 2005;18(2):91–94. doi: 10.1087/0953151053584984

[19] Murata T, Saito N, Koyama E, et al. Data-driven analysis and visualization of dielectric properties curated from scientific literature. Sci Technol Adv Mater Methods [Internet]. 2025;5(1). doi: 10.1080/27660400.2025.2485018

[20] Fujita E, Liu C, Ishikawa A, et al. Comprehensive experimental datasets of quasicrystals and their approximants. Sci Data. 2024;11(1):1–9. doi: 10.1038/s41597-024-04043-z

[21] Yamashita Y, Yagi T, Baba T. Development of network database system for thermophysical property data of thin films. Jpn J Appl Phys. 2011;50(11S):11RH03. doi: 10.1143/JJAP.50.11RH03

[22] Brinson LC, Deagen M, Chen W, et al. Polymer nanocomposite data: curation, frameworks, access, and potential for discovery and design. ACS Macro Lett. 2020;9(8):1086–1094. doi: 10.1021/acsmacrolett.0c00264

[23] Jacobsson TJ, Hultqvist A, García-Fernández A, et al. An open-access database and analysis tool for perovskite solar cells based on the FAIR data principles. Nat Energy. 2021;7(1):107–115. doi: 10.1038/s41560-021-00941-3

[24] Zakutayev A, Wunder N, Schwarting M, et al. An open experimental database for exploring inorganic materials. Sci Data. 2018;5(1):180053. doi: 10.1038/sdata.2018.53

[25] Marin F, Rohatgi A, Charlot S. WebPlotDigitizer, a polyvalent and free software to extract spectra from old astronomical publications: application to ultraviolet spectropolarimetry [internet]. ArXiv [astro-ph.IM]. 2017. Available from: http://arxiv.org/abs/1708.02025

[26] Liu F, Eisenschlos JM, Piccinno F, et al. Deplot: one-shot visual language reasoning by plot-to-table translation [internet]. ArXiv [cs.cl]. 2022 [cited 2025 Aug 28]. Available from: http://arxiv.org/abs/2212.10505

[27] Lee J, Lee W, Kim J. MatGD: materials graph digitizer. ACS Appl Mater Interface. 2024;16(1):723–730. doi: 10.1021/acsami.3c14781

[28] Mavračić J, Court CJ, Isazawa T, et al. Chemdataextractor 2.0: autopopulated ontologies for materials science. J Chem Inf Model. 2021;61(9):4280–4289. doi: 10.1021/acs.jcim.1c00446

[29] Snyder GJ, Toberer ES. Complex thermoelectric materials. Nat Mater. 2008;7(2):105–114. doi: 10.1038/nmat2090

[30] Gaultois MW, Sparks TD, Borg CKH, et al. Data-driven review of thermoelectric materials: performance and resource considerations. Chem Mater. 2013;25(15):2911–2920. doi: 10.1021/cm400893e
[31] Gaultois MW, Oliynyk AO, Mar A, et al. Perspective: web-based machine learning models for real-time screening of thermoelectric materials properties. APL Mater [Internet]. 2016;4(5). doi: 10.1063/1.4952607
[32] Mbaye MT, Pradhan SK, Bahoura M. Data-driven thermoelectric modeling: current challenges and prospects. J Appl Phys [Internet]. 2021;130(19). doi: 10.1063/5.0054532
[33] Antunes LM, Vikram PJ, Powell AV, et al. Machine learning approaches for accelerating the discovery of thermoelectric materials. In: An Y, editor. Machine learning in materials informatics: methods and applications. Washington (DC): American Chemical Society; 2022. p. 1–32.
[34] Wang X, Sheng Y, Ning J, et al. A critical review of machine learning techniques on thermoelectric materials. J Phys Chem Lett. 2023;14(7):1808–1822. doi: 10.1021/acs.jpclett.2c03073
[35] Yildirim E, Yelgel Ö C. Using machine learning techniques to discover novel thermoelectric materials. In: Ismail D, editor. New materials and devices for thermoelectric power generation [working title]. London, (UK): IntechOpen; 2023. p. 36–67.
[36] Barua NK, Lee S, Oliynyk AO, et al. Recent strides in artificial intelligence for predicting thermoelectric properties and materials discovery. J Phys Energy. 2025;7(2):021001. doi: 10.1088/2515-7655/adba87
[37] Yelgel ÖC, Yelgel C. A review of machine learning approaches for the discovery of thermoelectric materials. Adv Phys X [Internet]. 2025;10(1). doi: 10.1080/23746149.2025.2536269
[38] Zaki M, Jayadeva M. Mascqa: investigating materials science knowledge of large language models. Digit Discov. 2024;3(2):313–327.
[39] Kang Y, Kim J. ChatMOF: an artificial intelligence system for predicting and generating metal-organic frameworks using large language models. Nat Commun. 2024;15(1):4705. doi: 10.1038/s41467-024-48998-4
[40] Livne M, Miftahutdinov Z, Tutubalina E, et al. Nach0: multimodal natural and chemical languages foundation model. Chem Sci. 2024;15(22):8380–8389. doi: 10.1039/D4SC00966E
[41] Jia S, Zhang C, Fung V. Llmatdesign: autonomous materials discovery with large language models [internet]. Arxiv [cond-mat.mtrl-sci]. 2024. Available from: http://arxiv.org/abs/2406.13163
[42] Zhang H, Song Y, Hou Z, et al. HoneyComb: a flexible LLM-based agent system for materials science [Internet]. arXiv [cs.CL]. 2024. Available from: http://arxiv.org/abs/2409.00135
[43] Ni Z, Li Y, Hu K, et al. MatPilot: an LLM-enabled AI materials scientist under the framework of human-machine collaboration [Internet]. arXiv [physics.soc-ph]. 2024 [cited 2025 Aug 28]. Available from: http://arxiv.org/abs/2411.08063
[44] Zhang Y, Itani S, Khanal K, et al. Gptarticleextractor: an automated workflow for magnetic material database construction. J Magn Magn Mater. 2024;597 (172001):172001. doi: 10.1016/j.jmmm.2024.172001
[45] Yi GH, Choi J, Song H, et al. MaTableGPT: gPT-based table data extractor from materials science literature. Adv Sci (Weinh). 2025;12(16):e2408221. doi: 10.1002/advs.202408221
[46] Tang Y, Xu W, Cao J, et al. MatterChat: a multi-modal LLM for material science [internet]. arXiv [cs.AI]. 2025. Available from: http://arxiv.org/abs/2502.13107
[47] Itani S, Zhang Y, Zang J. Large language model-driven database for thermoelectric materials [Internet]. arXiv [cond-mat.mtrl-sci]. 2024 [cited 2025 Jan 7]. Available from: http://arxiv.org/abs/2501.00564
[48] Itani S, Zhang Y, Zang J. Northeast materials database (NEMAD): enabling discovery of high transition temperature magnetic compounds [Internet]. arXiv [cond-mat.mtrl-sci]. 2024 [cited 2025 Jan 23]. Available from: http://arxiv.org/abs/2409.15675
[49] Foppiano L, Dieb S, Suzuki A, et al. Supermat: construction of a linked annotated dataset from superconductors-related publications. Sci Technol Adv Mater Methods. 2021;1(1):34–44. doi: 10.1080/27660400.2021.1918396
[50] Compton M, Barnaghi P, Bermudez L, et al. The SSN ontology of the W3C semantic sensor network incubator group. Web Semant. 2012;17:25–32. doi: 10.1016/j.websem.2012.05.003
[51] Zhang X, Zhao C, Wang X. A survey on knowledge representation in materials science and engineering: an ontological perspective. Comput Ind. 2015;73:8–22. doi: 10.1016/j.compind.2015.07.005
[52] Takahashi L, Miyazato I, Takahashi K. Redesigning the materials and catalysts database construction process using ontologies. J Chem Inf Model. 2018;58 (9):1742–1754. doi: 10.1021/acs.jcim.8b00165
[53] Ishii M, Sakamoto K. Structuring superconductor data with ontology: reproducing historical datasets as knowledge bases. Sci Technol Adv Mater Methods. 2023;3(1):2223051. doi: 10.1080/27660400.2023.2223051
[54] American Chemical Society. Terms and conditions [Internet]. Washington, DC: American Chemical Society; [date unknown; cited 2025 Apr 15].
[55] Elsevier. Terms and conditions [Internet]. Amsterdam: Elsevier; 2025 [cited 2025 Apr 15; updated 2025 Jun 10]. Available from: https://www.elsevier.com/legal/elsevier-website-terms-and-conditions
[56] Wiley. Terms and conditions of use [Internet]. Hoboken (NJ): John Wiley & Sons, Inc; [cited 2025 Apr 15]. Available from: https://www.wiley.com/en-us/terms-of-use
[57] Hendricks G, Tkaczyk D, Lin J, et al. Crossref: the sustainable source of community-owned scholarly metadata. Quant Sci Stud. 2020;1(1):414–427. doi: 10.1162/qss_a_00022
[58] Anthropic. The Claude 3 model family: Opus, Sonnet, Haiku [Internet]. San Francisco (CA): Anthropic Ltd.; 2024 Mar 4 [cited 2025 Apr 15]. Available from: https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf
[59] Garmroudi F, Parzer M, Riss A, et al. Solubility limit and annealing effects on the microstructure & thermoelectric properties of Fe2V1−xTaxAl1−ySiy Heusler compounds. Acta Mater. 2021;212 (116867):116867.