

# Submissions and Reflections from the 2024 Large Language Model (LLM) Hackathon for Applications in Materials Science and Chemistry

Yoel Zimmermann<sup>†</sup><sup>38</sup>, Adib Bazgir<sup>†</sup><sup>54</sup>, Zartashia Afzal<sup>1</sup>, Fariha Agbere<sup>2</sup>, Qianxiang Ai<sup>3</sup>, Nawaf Alampara<sup>4</sup>, Alexander Al-Feghali<sup>5</sup>, Mehrad Ansari<sup>6</sup>, Dmytro Antypov<sup>7</sup>, Amro Aswad<sup>8</sup>, Jiaru Bai<sup>9</sup>, Viktoriia Baibakova<sup>10</sup>, Devi Dutta Biswajeet<sup>11</sup>, Erik Bitzek<sup>12</sup>, Joshua D. Bocarsly<sup>13</sup>, Anna Borisova<sup>13</sup>, Andres M. Bran<sup>13</sup>, L. Catherine Brinson<sup>17</sup>, Marcel Moran Calderon<sup>13</sup>, Alessandro Canalicchio<sup>14</sup>, Victor Chen<sup>15</sup>, Yuan Chiang<sup>10,16</sup>, Defne Cincio<sup>17</sup>, Benjamin Charmes<sup>9</sup>, Vikrant Chaudhary<sup>18,19</sup>, Zizhang Chen<sup>20</sup>, Min-Hsueh Chiu<sup>21</sup>, Judith Clymo<sup>7</sup>, Kedar Dabhadkar<sup>22</sup>, Nathan Daelman<sup>18</sup>, Archit Datar<sup>14</sup>, Wibe A. de Jong<sup>16</sup>, Matthew L. Evans<sup>23,24</sup>, Maryam Ghazizade Fard<sup>25</sup>, Giuseppe Fisicaro<sup>26</sup>, Abhijeet Sadashiv Gangan<sup>27</sup>, Janine George<sup>4,43</sup>, Jose D. Cojal Gonzalez<sup>28</sup>, Michael Götte<sup>28</sup>, Ankur K. Gupta<sup>16</sup>, Hassan Harb<sup>20</sup>, Pengyu Hong<sup>21</sup>, Abdelrahman Ibrahim<sup>4</sup>, Ahmed Ilyas<sup>18</sup>, Alishba Imran<sup>31</sup>, Kevin Ishimwe<sup>2</sup>, Ramsey Issa<sup>33</sup>, Kevin Maik Jablonka<sup>4</sup>, Colin Jones<sup>2</sup>, Tyler R. Josephson<sup>2</sup>, Gergely Juhasz<sup>34</sup>, Sarthak Kapoor<sup>18</sup>, Rongda Kang<sup>35</sup>, Ghazal Khalighinejad<sup>17</sup>, Sartaj Takrim Khan<sup>18</sup>, Sascha Klawohn<sup>18</sup>, Suneel Kuman<sup>36</sup>, Alvin Noe Ladines<sup>18</sup>, Sarom Leang<sup>37</sup>, Magdalena Lederbauer<sup>13,38</sup>, Sheng-Lun (Mark) Liao<sup>35</sup>, Hao Liu<sup>39</sup>, Xuefeng Liu<sup>15,73</sup>, Stanley Loo<sup>8</sup>, Sandeep Madireddy<sup>20</sup>, Piyush Ranjan Maharana<sup>72</sup>, Shagun Maheshwari<sup>40</sup>, Soroush Mahjoubi<sup>3</sup>, Jos'e A. Márquez<sup>18</sup>, Rob Mills<sup>13</sup>, Trupti Mohanty<sup>33</sup>, Bernadette Mohr<sup>18,41</sup>, Seyed Mohamad Moosavi<sup>6,8</sup>, Alexander Moßhammer<sup>14</sup>, Amirhossein D. Naghdipour<sup>42</sup>, Aakash Naik<sup>4,43</sup>, Oleksandr Narykov<sup>20</sup>, Hampus Näsström<sup>18</sup>, Xuan Vu Nguyen<sup>44</sup>, Xinyi Ni<sup>30</sup>, Dana O'Connor<sup>45</sup>, Teslim Olayiwola<sup>46</sup>, Federico Ottomano<sup>7</sup>, Aleyna Beste Ozhan<sup>3</sup>, Sebastian Pagel<sup>47</sup>, Chiku Parida<sup>48</sup>, Jaehee Park<sup>15</sup>, Vraj Patel<sup>12</sup>, Elena Patyukova<sup>7</sup>, Martin Hoffmann Petersen<sup>48</sup>, Luis Pinto<sup>49</sup>, Jos'e M. Pizarro<sup>18</sup>, Dieter Plessers<sup>50</sup>, Tapashree Pradhan<sup>50</sup>, Utkarsh Pratiush<sup>51</sup>, Charishma Puli<sup>2</sup>, Andrew Qin<sup>15</sup>, Mahyar Rajabi<sup>8</sup>, Francesco Ricci<sup>16</sup>, Elliot Risch<sup>52</sup>, Martíno Ríos-García<sup>4,53</sup>, Aritra Roy<sup>71</sup>, Tehseen Rugh<sup>14</sup>, Hasan M Sayeed<sup>33</sup>, Markus Scheidgen<sup>18</sup>, Mara Schilling-Wilhelmi<sup>4</sup>, Marcel Schloz<sup>18</sup>, Fabian Schöppach<sup>18</sup>, Julia Schumann<sup>18</sup>, Philippe Schwaller<sup>13</sup>, Marcus Schwarting<sup>15</sup>, Samiha Sharlin<sup>2</sup>, Kevin Shen<sup>55</sup>, Jiale Shi<sup>3</sup>, Pradip Si<sup>56</sup>, Jennifer D'Souza<sup>57</sup>, Taylor Sparks<sup>33</sup>, Suraj Sudhakar<sup>15</sup>, Leopold Talirz<sup>32</sup>, Dandan Tang<sup>58</sup>, Olga Taran<sup>59</sup>, Carla Terboven<sup>28</sup>, Mark Tropin<sup>61</sup>, Anastasiia Tsymbal<sup>62,63</sup>, Katharina Ueltzen<sup>43</sup>, Pablo Andres Unzueta<sup>64</sup>, Archit Vasan<sup>20</sup>, Tirtha Vinchurkar<sup>40</sup>, Trung Vo<sup>11</sup>, Gabriel Vogel<sup>65</sup>, Christoph Völker<sup>14</sup>, Jan Weinreich<sup>66</sup>, Faradawn Yang<sup>5</sup>, Mohd Zaki<sup>67</sup>, Chi Zhang<sup>7</sup>, Sylvester Zhang<sup>5</sup>, Weijie Zhang<sup>58</sup>, Ruijie Zhu<sup>15</sup>, Shang Zhu<sup>69</sup>, Jan Janssen<sup>70</sup>, Calvin Li<sup>75</sup>, Ian Foster<sup>15,20</sup>, and Ben Blaiszik<sup>\*15,20</sup>

<sup>1</sup>University of the Punjab

<sup>2</sup>University of Maryland, Baltimore County

<sup>3</sup>Massachusetts Institute of Technology

<sup>4</sup>Friedrich-Schiller-Universität Jena

<sup>5</sup>McGill University  
<sup>6</sup>Acceleration Consortium  
<sup>7</sup>University of Liverpool  
<sup>8</sup>University of Toronto  
<sup>9</sup>University of Cambridge  
<sup>10</sup>University of California at Berkeley  
<sup>11</sup>University of Illinois at Chicago  
<sup>12</sup>University of Houston  
<sup>13</sup>EPFL  
<sup>14</sup>iteratec GmbH  
<sup>15</sup>University of Chicago  
<sup>16</sup>Lawrence Berkeley National Laboratory  
<sup>17</sup>Duke University  
<sup>18</sup>Humboldt University of Berlin  
<sup>19</sup>Technology University of Darmstadt  
<sup>20</sup>Argonne National Laboratory  
<sup>21</sup>University of Southern California  
<sup>22</sup>Lam Research  
<sup>23</sup>Université catholique de Louvain  
<sup>24</sup>Matgenix SRL  
<sup>25</sup>Queen's University  
<sup>26</sup>CNR Institute for Microelectronics and Microsystems, Catania  
<sup>27</sup>University of California at Los Angeles  
<sup>28</sup>Helmholtz-Zentrum Berlin für Materialien und Energie GmbH  
<sup>30</sup>Brandeis University  
<sup>31</sup>Kleiner Perkins  
<sup>32</sup>Schott  
<sup>33</sup>University of Utah  
<sup>34</sup>Tokyo Institute of Technology  
<sup>35</sup>Factorial Energy  
<sup>36</sup>Molecular Forecaster  
<sup>37</sup>EP Analytics, Inc.  
<sup>38</sup>ETH Zurich  
<sup>39</sup>Fordham University  
<sup>40</sup>Carnegie Mellon University  
<sup>41</sup>University of Amsterdam  
<sup>42</sup>IDEAS NCBR  
<sup>43</sup>Federal Institute of Materials Research and Testing (BAM)  
<sup>44</sup>Università degli Studi di Milano  
<sup>45</sup>Pittsburgh Supercomputing Center  
<sup>46</sup>Louisiana State University  
<sup>47</sup>University of Glasgow  
<sup>48</sup>Technical University of Denmark  
<sup>49</sup>Independent Researcher  
<sup>50</sup>KU Leuven  
<sup>51</sup>University of Tennessee, Knoxville

<sup>52</sup>Enterprise Knowledge  
<sup>53</sup>Instituto de Ciencia y Tecnología del Carbono  
<sup>54</sup>University of Missouri-Columbia  
<sup>55</sup>NobleAI  
<sup>56</sup>University of North Texas  
<sup>57</sup>TIB Leibniz Information Centre for Science and Technology  
<sup>58</sup>University of Virginia  
<sup>59</sup>University of California at Davis  
<sup>60</sup>Helmholtz-Zentrum Berlin für Materialien und Energie GmbH  
<sup>61</sup>Windmill Labs  
<sup>62</sup>Rutgers University  
<sup>63</sup>University of Pennsylvania  
<sup>64</sup>Stanford University  
<sup>65</sup>Delft University of Technology  
<sup>66</sup>Quastify GmbH  
<sup>67</sup>Indian Institute of Technology Delhi  
<sup>68</sup>RWTH Aachen University  
<sup>69</sup>University of Michigan-Ann Arbor  
<sup>70</sup>Max-Planck Institute for Sustainable Materials  
<sup>71</sup>London South Bank University  
<sup>72</sup>CSIR-National Chemical Laboratory  
<sup>73</sup>LinkDot.AI  
<sup>74</sup>Celanese Corporation  
<sup>75</sup>Fum Technologies, Inc.

\*Corresponding author: blaiszik@uchicago.edu These authors also contributed substantially to compiling team results and other paper writing tasks

### Abstract

Here, we present the outcomes from the second Large Language Model (LLM) Hackathon for Applications in Materials Science and Chemistry, which engaged participants across global hybrid locations, resulting in 34 team submissions. The submissions spanned seven key application areas and demonstrated the diverse utility of LLMs for applications in (1) molecular and material property prediction; (2) molecular and material design; (3) automation and novel interfaces; (4) scientific communication and education; (5) research data management and automation; (6) hypothesis generation and evaluation; and (7) knowledge extraction and reasoning from scientific literature. Each team submission is presented in a summary table with links to the code and as brief papers in the appendix. Beyond team results, we discuss the hackathon event and its hybrid format, which included physical hubs in Toronto, Montreal, San Francisco, Berlin, Lausanne, and Tokyo, alongside a global online hub to enable local and virtual collaboration. Overall, the event highlighted significant improvements in LLM capabilities since the previous year's hackathon, suggesting continued expansion of LLMs for applications in materials science and chemistry research. These outcomes demonstrate the dual utility of LLMs as both multipurpose models for diverse machine learning tasks and platforms for rapid prototyping custom applications in scientific research.

## Introduction

Science hackathons have emerged as a powerful tool for fostering collaboration building, innovation, and rapid problem-solving in the scientific community [1–3]. By leveraging social media, virtual platforms, and

hybrid event structures, such hackathons can be organized in a cost-effective manner while maximizing their impact and reach. In this article, we first introduce the project submissions to the second Large Language Model Hackathon for Applications in Materials Science and Chemistry, detailing the broad classes of problems addressed by teams, while analyzing trends and patterns in the approaches taken. We then present each team submission in turn, plus a summary table with the names of team members and links to code repositories where available. Finally, we include the detailed project documents submitted by each team, showcasing the depth and breadth of innovation demonstrated during the hackathon.

## Overview of Submissions

The hackathon resulted in 34 team submissions (with 32 submissions with a written description included here), categorized as shown in [Table 1](#). From these submissions, we identified seven key application areas:

1. **Molecular and MaterialProperty Prediction**: Forecasting chemical and physical properties of molecules and materials using LLMs, particularly excelling in low-data environments and combining structured/unstructured data.
2. **Molecular and MaterialDesign**: Generation and optimization of novel molecules and materials using LLMs, including peptides, metal-organic frameworks, and sustainable construction materials.
3. **Automation and Novel Interfaces**: Development of natural language interfaces and automated workflows to simplify complex scientific tasks, making advanced tools and techniques more accessible to researchers.
4. **Scientific Communication and Education**: Enhancement of academic communication, automation of educational content creation, and facilitation of learning in materials science and chemistry.
5. **Research Data Management and Automation**: Streamlining the handling, organization, and processing of scientific data through LLM-powered tools and multimodal agents.
6. **Hypothesis Generation and Evaluation**: Generation, assessment, and validation of scientific hypotheses using LLMs, often combining multiple AI agents and statistical approaches.
7. **Knowledge Extraction and Reasoning**: Extraction of structured information from scientific literature and sophisticated reasoning about chemical and materials science concepts through knowledge graphs and multimodal approaches.

We next discuss each application area in more detail and highlight exemplar projects in each.

### 1. Molecular and Material Property Prediction

LLMs have rapidly advanced in this area, employing both textual and numerical data to forecast a wide range of properties. Recent studies [4–6] show LLMs performing comparably to, or even surpassing, conventional machine learning methods in this domain, *particularly in low-data environments*. Their flexibility in processing both structured and unstructured data, as well as their general applicability to regression tasks [7], makes them a powerful tool for diverse predictive tasks in molecular and materials science.

**Exemplar projects** The **Learning LOBSTERs** team (Ueltzen et al.) integrated bonding analysis data into LLMs to predict the phonon density of states (DOS) for crystal structures, showing that combining covalent bonding information with LLM capabilities can improve prediction accuracy. The **Liverpool Materials** team (Ottomano et al.) demonstrated how adding context from materials literature, via models like MatSciBert [8], could improve the prediction of Li-ion conductivity, particularly when dealing with limited training data. The **MolFoundation** team (Harb et al.) benchmarked ChemBERTa [9] and T5-Chem [10] for molecular property predictions, finding that pre-trained models performed comparably to fine-tuned models, suggesting that expensive fine-tuning might not always be required. Another contribution came from the **Geometric Geniuses** team (Weinreich et al.), who focused on incorporating 3D molecular geometries into LLMs, experimenting with encoding geometric features to improve prediction outcomes.

## 2. Molecular and Material Design

LLMs have also advanced in molecular and material design, proving capable in both settings [11–15], especially if pre-trained or fine-tuned with domain-specific data [16]. However, despite these advancements, LLMs still face limitations in practical applications [17].

**Exemplar projects:** During the hackathon, teams tackled these challenges through different approaches. The team behind **MC-Peptide** (Bran et al.) developed a workflow harnessing LLMs for the design of macrocyclic peptides (MCPs). By employing semantic search and constraint-based generation, they automated the extraction of literature data to propose MCPs with improved permeability, crucial for drug development [18, 19]. Meanwhile, the **MOF Innovators** team (Ansari et al.) focused on metal-organic frameworks (MOFs). Their AI agent utilized retrieval-augmented generation (RAG) [20] to incorporate design rules extracted from the literature, and an ensemble of fine-tuned surrogate models to optimize MOF's band gap property. In pursuit of sustainable materials, the **Green Construct** team (Canalicchio et al.) investigated small-scale LLMs such as Llama 3 [21] and Phi-3 [22] to streamline the design of alkali-activated concrete. Their models provided insights into reducing emissions of traditional materials like Portland cement through zero-shot learning.

## 3. Automation and Novel Interfaces

LLMs are increasingly used to simplify and streamline access to complex tools [23–25], autonomously plan and execute tasks [26], and interface with robotic systems in lab settings [27, 28]. These capabilities enhance the efficiency of scientific workflows, allowing researchers to focus on higher-level problem-solving rather than routine tasks. As LLMs continue to evolve, they are expected to further transform laboratory practices and democratize access to advanced experimental and computational techniques.

**Exemplar projects:** The team behind **LangSim** (Chiang et al.) addressed the complexity of atomistic simulation software by creating a natural language interface to automate the calculation of material properties such as bulk modulus. By integrating LangChain [29] agents with LLMs, LangSim enables flexible construction and execution of simulation workflows, reducing user barriers and expanding functionality to multi-component alloys. Similarly, **LLMScopilot** (Schloz & Gonzalez) seeks to simplify the operation of sophisticated microscopes by using LLM-powered agents to automate tasks like experimental parameter estimation. This could reduce the need for highly trained operators, making advanced microscopy tools more accessible. In the realm of Density Functional Theory (DFT) calculations, the team behind **T2DlLama** (Parida & Petersen) developed a tool that uses RAG to extract optimized simulation parameters from scientific literature, aiming to assist experimentalists with complex DFT setups. The tool simplifies the process of obtaining reliable parameters and reduces dependency on computational chemistry expertise. Complementing this, **Materials Agent** (Datar et al.) provides a tool-calling capability for cheminformatics, combining molecular property calculations, simulations, and document interaction through a natural language interface. This agent was developed to lower the barrier for researchers in cheminformatics and to accelerate the pace of research by integrating tools like RDKit [30] and custom simulations into an intuitive system.

## 4. Scientific Communication and Education

LLMs are transforming how scientific and educational content is created and shared, enhancing accessibility and personalized learning [31–34]. By automating tasks like question generation, feedback, and grading, LLMs streamline educational processes, freeing educators to focus on individual learning needs. Additionally, LLMs assist in translating complex scientific findings into accessible formats, broadening public engagement [34]. However, technological readiness, transparency, and ethical concerns around data privacy and bias remain critical challenges to address [31, 33].

**Exemplar projects:** One submission in this category is **MaSTeA** (Circi et al.), an automated system focused on evaluating the effectiveness of LLMs as teaching assistants by testing their performance on materials science questions from the MaScQA [35] dataset. By analyzing different types of questions—such as multiple-choice, match-the-following, and numerical problems—across 14 topical categories, the team identified the strengths and weaknesses of various models. An interactive platform was developed to facilitate easy model testing and, as models get better, to eventually help students practice answering questions and

understand solution steps. Meanwhile, the **LLMy Way** team (Zhu et al.) focused on simplifying the process of creating academic presentations by using LLMs to automatically generate structured slide decks from research articles. The tool formats summaries of papers into typical sections—such as background, methods, and results—and transforms them into presentation slides. It offers customization based on audience expertise and time constraints, aiming to make the communication of complex topics more efficient and effective. Lastly, the **WaterLLM** team (Baibakova et al.) sought to address water pollution challenges, particularly in decentralized communities lacking centralized water treatment infrastructure. They developed a chatbot that uses LLMs enhanced with RAG to suggest optimal water purification methods for microplastic contamination. Grounded in up-to-date scientific literature, the chatbot provides tailored purification protocols based on contaminant characteristics, resource availability, and cost considerations, promoting effective water treatment solutions for underserved areas.

## 5. Research Data Management and Automation

Various submissions were received in this area that attempt to enhance the management, accessibility, and automation of scientific data workflows using LLMs. These efforts, often leveraging multimodal agents, aim to simplify complex data handling, improve reproducibility, and accelerate insights across diverse scientific disciplines.

**Exemplar projects** **yeLLowhaMmer** (Evans et al.), a multimodal LLM-based data management agent automates data handling within electronic lab notebooks (ELNs) and laboratory information management systems (LIMS). The tool processes text and image instructions to generate Python code targeting the API of the *datalab* ELN/LIMS system [1], performing tasks such as summarizing experiments or digitizing handwritten notes. Powered by the Claude 3 Haiku model [36] and LangChain [29], **yeLLowhaMmer** was developed to simplify the interaction with scientific data and has future potential for integrating audiovisual data processing. In parallel, the team behind **LLMads** (Kapoor et al.) explored how LLMs can be used to automate data parsing by converting raw scientific data, like X-ray diffraction (XRD) measurements, into structured metadata schemas. **LLMads** aims to replace the need for custom parsers with models like Mixtral-8x7b [37], which extract data from raw files and populate schemas automatically. The process involves chunking raw data and prompt engineering, although challenges such as hallucinations and the extraction of multi-dimensional data still require further refinement. The **NOMAD Query Reporter** team (Daelman et al.) developed an LLM-based agent that uses RAG to generate context-aware summaries from large materials science repositories like NOMAD [38]. The tool produces detailed reports on experimental methodologies and results, which could help researchers in drafting publication sections. Using a self-hosted Llama3 70B model, the system engages in multi-turn conversations to ensure context retention. Further, in the **Speech-schema-filling** project, Nistörm et al. integrated speech recognition and LLMs to automate the conversion of spoken language into structured data within lab environments. This method, utilizing OpenAI's Whisper [39] for transcription and LLMs for schema selection and population, allows researchers to input experimental data verbally, which is then structured into JSON or Pydantic models for entry into ELNs, which could be especially helpful in situations where manual entry is difficult.

## 6. Hypothesis Generation and Evaluation

LLMs can be leveraged to streamline scientific inquiry, hypothesis generation, and verification. Recent work across psychology, astronomy, and biomedical research demonstrates their capacity to generate novel, validated hypotheses by integrating domain-specific data structures like causal graphs [40–43]. Although still largely untapped in chemistry and materials science, this approach holds substantial promise for accelerating discovery and innovation in these fields [44, 45].

**Exemplar projects** In an individual contribution, **Marcus Schwarting** employed LLMs and temporal Bayesian statistics to assess scientific claims, exemplified by evaluating the hypothesis that LK-99 is a room-temperature superconductor. By using LLM-based natural language inference (NLI) to classify relevant studies and applying Bayesian updating, his system tracks how the scientific consensus evolves over time, with the goal of allowing even non-experts to gauge quickly the validity of a claim. In a related project, the **Thoughtful Beavers** team (Ozhan & Mahjoubi) prototyped a Multi-Agent System of specialized LLMs that aims to accelerate scientific hypothesis generation by coordinating agents that extract background

information, generate inspirations, and propose hypotheses, which are then evaluated for feasibility, utility, and novelty. By applying the “Tree of Thoughts” framework [46], this system streamlines the creative process and improves the quality of scientific hypotheses, as demonstrated through a case study on sustainable concrete design. Another project, **ActiveScience** (Chiu) integrated LLMs, knowledge graphs, and RAG to extract high-level insights from scientific articles. Focused on materials science research, particularly alloys, this framework organizes extracted information into a Neo4j knowledge graph, allowing for complex querying and discovery. Additionally, a first-pass peer review system was developed using fine-tuned LLMs, **G-Peer-T** (Al-Feghali & Zhang), to assess materials science papers. By analyzing the log probabilities of abstracts, the system flags works that deviate from typical scientific language patterns, helping to identify both highly innovative and potentially nonsensical research. Preliminary findings suggest that highly cited papers tend to use less typical language, highlighting the potential for LLMs to support the peer review process and detect outlier research.

## 7. Knowledge Extraction and Reasoning

Here the interest is in LLMs enabling extraction of structured scientific knowledge from unstructured text, assisting researchers in navigating complex academic content [47–50]. These systems streamline tasks like named entity recognition and relation extraction, offering flexible solutions tailored to materials science and chemistry [48]. Tool-augmented frameworks help LLMs address complex reasoning by leveraging scientific tools and resources, expanding their utility as assistants in scientific research [51].

**Exemplar projects** The team behind **ChemQA** (Khalighinejad et al.) created a multimodal question-and-answering dataset for chemical understanding. It highlights the importance of combining textual and visual inputs (e.g., SMILES and molecule images) to improve model accuracy in tasks like counting atoms, molecular weight calculations, and retrosynthesis planning. This multimodal approach shows how foundational models in chemistry benefit from rich, diverse data representations. In the realm of lithium metal batteries, **LithiumMind** (Ni et al.), integrated LLMs to extract Coulombic Efficiency (CE) and electrolyte information from scientific literature. The team also developed a chatbot powered by RAG to answer queries on lithium battery research. By constructing a knowledge graph, the system visualizes relationships between materials and properties, enhancing research accessibility and fostering domain-specific insights. The **KnowMat** project (Sayeed et al.) tackled the transformation of unstructured materials science literature into structured formats, using LLMs to extract essential information and convert it into machine-readable data (JSON). KnowMat’s customizable prompts and integration capabilities streamline data extraction and analysis, making it a potentially valuable tool for researchers looking to accelerate their workflow in materials science. A similar project, **Ontosynthesis** (Ait et al.), explored extracting structured data from organic synthesis texts using LLMs. By employing Resource Description Framework (RDF) graphs, the project improved how chemical reactions are represented, with the extracted data aiding in retrosynthesis and condition recommendation tasks. This approach could help bridge the gap between unstructured chemical descriptions and standardized data formats. For high entropy alloys (HEAs) in hydrogen storage, the **Insight Engineers** team (Pradhan & Biswajeet) explored synthetic data generation as a means to accelerate predictive modeling. The team proposed using LLMs like GPT-4 [52], alongside custom prompts and a RAG framework, to generate synthetic data for machine learning interatomic potentials (MLIPs) to overcome the computational challenges associated with HEA properties prediction. Lastly, **GlossaGen** (Lederbauer et al.) tackled the challenge of generating glossaries for academic papers and grant proposals. It uses LLMs to extract scientific terms and definitions from PDF or TeX files and presents them in a knowledge graph. This visualization of relationships between concepts enhances understanding, and there are possible expansions of the tool to offer LaTeX integration and more refined ontology creation to aid researchers in navigating complex scientific terminology.

## Hackathon Event Overview

On May 9th, 2024, we hosted the second annual Large Language Model (LLM) Hackathon for Applications in Materials Science and Chemistry. The event brought together students, postdocs, researchers from industry, citizen scientists, and more, with participants joining both virtually and at in-person sites across

# LLM HACKATHON FOR APPLICATIONS IN MATERIALS AND CHEMISTRY

- 1 ONLINE HUB FOR WORLDWIDE ACCESS
- 7 IN-PERSON LOCATIONS



Figure 1: LLM Hackathon for Applications in Materials and Chemistry hybrid hackathon. Researchers were able to participate from both remote and in-person locations (purple pins).

multiple continents (Figure 1). The event was a follow-up to the previous hackathon described in detail here. [53] The event began with a kickoff panel featuring leading researchers from academia and industry, including Elsa Olivetti (MIT), Jon Reifsneider (Duke), Michael Craig (Valence Laboratories), and Marwin Segler (Microsoft). The charge of the hackathon was intentionally open-ended; i.e., to explore the vast potential application space, and create tangible demonstrations of the most innovative, impactful, and scalable solutions in a constrained time using open-source and best-in-class multimodal models applied to problems in materials science and chemistry.

Event registration included 556 participants and over 120 researchers comprising 34 teams that submitted completed projects. The hybrid format proved particularly successful, with physical hub locations in Toronto, Montreal, San Francisco, Berlin, Lausanne, and Tokyo facilitating local collaboration while maintaining global connectivity across the physical hubs and remote participants through virtual platforms (Figure 1). This distributed approach enabled researchers to participate from either a local site or remotely from anywhere on Earth. This format blended the strengths of in-person events with the flexibility of remote participation leading to an inclusive event that led to team formation that crossed international and institutional boundaries, the submitted projects in this paper, and the growth of a new persistent online community of 483 researchers via Slack.

## Conclusion

The LLM Hackathon for Applications in Materials Science and Chemistry has demonstrated the dual utility and immense promise of large language models serving as 1) multipurpose models for diverse machine learning tasks and 2) platforms for rapid prototyping. Participants effectively utilized LLMs to tackle specific challenges while rapidly evaluating their ideas over a short 24-hour period, showcasing their ability to enhance the efficiency and creativity of research processes in highly diverse ways. It's important to note that many projects benefited from significant advancements in LLM performance since last year's hackathon. That is, the performance across the diverse application space was improved simply via the release of new versions of Gemini, ChatGPT, Claude, Llama, and other models. If this trend continues, we expect to see even broader applications in subsequent hackathons, and in materials science and chemistry more generally.

Additionally, the hackathon hybrid format has proven effective towards creating new scientific collaborations and communities. By uniting individuals from various backgrounds and areas of expertise, these events facilitate knowledge exchange and promote interdisciplinary approaches, which are essential for advancing research in this rapidly evolving field.

As the integration of LLMs continues to expand, collaborative initiatives like hackathons will play a critical role in driving innovation and addressing complex challenges in chemistry, materials science, and beyond. The outcomes from this event highlight the significance of leveraging LLMs for their adaptability and their potential to accelerate the development of new concepts and applications.

Table 1: Overview of the tools developed by the various tools, and links to source code repositories. Full descriptions of the projects can be found in the appendix.

Project	Authors	Links
<b>Molecular and Material Property Prediction</b>		
Leveraging Orbital-Based Bonding Analysis Information in LLMs	Katharina Ueltzen, Aakash Naik, Janine George	<a href="#">GitHub</a>
Context-Enhanced Material Property Prediction (CEMPP)	Federico Ottomano, Elena Patyukova, Judith Clymo, Dmytro Antypov, Chi Zhang, Aritra Roy, Piyush Ranjan Maharana, Weijie Zhang, Xuefeng Liu, Erik Bitzek	<a href="#">GitHub</a>
MolFoundation: Benchmarking Chemistry LLMs on Predictive Tasks	Hassan Harb, Xuefeng Liu, Anastasiia Tsybala, Oleksandr Narykov, Dana O'Connor, Shagun Maheshwari, Stanley Lo, Archit Vasan, Zartashia Afzal, Kevin Shen	<a href="#">GitHub</a>
3D Molecular Feature Vectors for Large Language Models	Jan Weinreich, Ankur K. Gupta, Amirhossein D. Naghdi, Alishba Imran	<a href="#">GitHub</a>
LLMSpectrometry	Tyler Josephson, Fariha Agbere, Kevin Ishimwe, Colin Jones, Charishma Puli, Samiha Sharlin, Hao Liu	<a href="#">GitHub</a>
<b>Molecular and Material Design</b>		
MC-Peptide: An Agentic Workflow for Data-Driven Design of Macrocyclic Peptides	Andres M. Bran, Anna Borisova, Marcel M. Calderon, Mark Tropin, Rob Mills, Philippe Schwaller	<a href="#">GitHub</a>

Project	Authors	Links
Leveraging AI Agents for Designing Low Band Gap Metal-Organic Frameworks	Sartaaj Khan, Mahyar Rajabi, Amro Aswad, Seyed Mohamad Moosavi, Mehrad Ansari	<a href="#">GitHub</a>
How Low Can You Go? Leveraging Small LLMs for Material Design	Alessandro Canalicchio, Alexander Moßhammer, Tehseen Rug, Christoph Völker	<a href="#">GitHub</a>
<b>Automation and Novel Interfaces</b>		
LangSim	Yuan Chiang, Giuseppe Fisicaro, Greg Juhasz, Sarom Leang, Bernadette Mohr, Utkarsh Pratiush, Francesco Ricci, Leopold Talirz, Pablo A. Unzueta, Trung Vo, Gabriel Vogel, Sebastian Pagel, Jan Janssen	<a href="#">GitHub</a>
LLMicroscopilot: assisting microscope operations through LLMs	Marcel Schloz, Jose C. Gonzalez	<a href="#">GitHub</a>
T2DlIlama: Harnessing Language Model for Density Functional Theory (DFT) Parameter Suggestion	Chiku Parida, Martin H. Petersen	<a href="#">GitHub</a>
Materials Agent: An LLM-Based Agent with Tool-Calling Capabilities for Cheminformatics	Archit Datar, Kedar Dabhadkar	<a href="#">GitHub</a>
LLM with Molecular Augmented Token	Luis Pinto, Xuan Vu Nguyen, Tirtha Vinchurkar, Pradip Si, Suneel Kuman	<a href="#">GitHub</a>
<b>Scientific Communication and Education</b>		
MaSTeA: Materials Science Teaching Assistant	Defne Circi, Abhijeet S. Gangan, Mohd Zaki	<a href="#">GitHub</a>
LLMy-Way	Ruijie Zhu, Faradawn Yang, Andrew Qin, Suraj Sudhakar, Jaehee Park, Victor Chen	<a href="#">GitHub</a>
WaterLLM: Creating a Custom ChatGPT for Water Purification Using PromptEngineering Techniques	Viktoria Baibakova, Maryam G. Fard, Teslim Olaiwola, Olga Taran	<a href="#">GitHub</a>
<b>Research Data Management and Automation</b>		
yeLLowhaMMer: A Multi-modal Tool-calling Agent for Accelerated Research Data Management	Matthew L. Evans, Benjamin Charmes, Vraj Patel, Joshua D. Bocarsly	<a href="#">GitHub</a>
LLMads	Sarthak Kapoor, Jos'e M. Pizarro, Ahmed Ilyas, Alvin N. Ladines, Vikrant Chaudhary	<a href="#">GitHub</a>
NOMAD Query Reporter: Automating Research Data Narratives	Nathan Daelman, Fabian Schöppach, Carla Terboven, Sascha Klawohn, Bernadette Mohr	<a href="#">GitHub</a>

Project	Authors	Links
Speech-schema-filling:Creating Structured Data Directly from Speech	Hampus N"asstr"om, Julia Schumann, Michael G"otte, Jos"e A. M"rquez	<a href="#">GitHub</a>
<b>Hypothesis Generation and Evaluation</b>		
Leveraging LLMs for Bayesian Temporal Evaluation of Scientific Hypotheses	Marcus Schwarting	<a href="#">GitHub</a>
Multi-Agent Hypothesis Generation and Verification through Tree of Thoughts and Retrieval Augmented Generation	Aleyna Beste Ozhan, Soroush Mahjoubi	<a href="#">GitHub</a>
ActiveScience	Min-Hsueh Chiu	<a href="#">GitHub</a>
G-Peer-T: LLM Probabilities For Assessing Scientific Novelty and Nonsense	Alexander Al-Feghali, Sylvester Zhang	<a href="#">GitHub</a>
<b>Knowledge Extraction and Reasoning</b>		
ChemQA	Ghazal Khalighinejad, Shang Zhu, Xuefeng Liu	<a href="#">GitHub</a>
LithiumMind - Leveraging Language Models for Understanding Battery Performance	Xinyi Ni, Zizhang Chen, Rongda Kang, Sheng-Lun Liao, Pengyu Hong, Sandeep Madireddy	<a href="#">GitHub</a>
KnowMat: Transforming Unstructured Material Science Literature into Structured Knowledge	Hasan M. Sayeed, Ramsey Issa, Trupti Mohanty, Taylor Sparks	<a href="#">GitHub</a>
Ontosynthesis	Qianxiang Ai, Jiaru Bai, Kevin Shen, Jennifer D'Souza, Elliot Risch	<a href="#">GitHub</a>
Knowledge Graph RAG for Polymer Simulation	Jiale Shi, Weijie Zhang, Dandan Tang, Chi Zhang	<a href="#">GitHub</a>
Synthetic Data Generation and Insightful Machine Learning for High Entropy Alloy Hydrides	Tapashree Pradhan, Devi Dutta Biswajeet	<a href="#">GitHub</a>
Chemsense:Are large language models aligned with human chemical preference?	Mart"no R"ios-Garc"ia, Nawaf Alampara, Mara Schilling-Wilhelmi, Abdelrahman Ibrahim, Kevin Maik Jablonka	<a href="#">GitHub</a>
GlossaGen	Magdalena Lederbauer, Dieter Plessers, Philippe Schwaller	<a href="#">GitHub</a>

## Acknowledgments

Planning for this event was supported by NSF Awards #2226419 and #220989. We would like to thank event sponsors who provided platform credits and prizes for teams, including RadicalAI, Iteratec, Reincarnate, Acceleration Consortium, and Neo4j.

## References

[1] A. Nolte, L. B. Hayden and J. D. Herbsleb, *Proc. ACM Hum.-Comput. Interact.*, 2020, **4**, 1–23. <https://doi.org/10.1145/3392830>

[2] E. P. P. Pe-Than and J. D. Herbsleb, in *Lecture Notes in Computer Science*, Springer, 2019, vol. 11546, pp. 27–37. [https://doi.org/10.1007/978-3-030-15742-5\\_3](https://doi.org/10.1007/978-3-030-15742-5_3)

[3] B. Heller, A. Amir, R. Waxman and Y. Maaravi, *J. Innov. Entrep.*, 2023, **12**, 1. <https://doi.org/10.1186/s13731-023-00269-0>

[4] K. M. Jablonka, P. Schwaller, A. Ortega-Guerrero and B. Smit, *Nat. Mach. Intell.*, 2024, **6**, 161–169. <https://doi.org/10.1038/s42256-023-00788-1>

[5] C. Qian, H. Tang, Z. Yang, H. Liang and Y. Liu, *arXiv*, 2023. <https://arxiv.org/abs/2307.07443>

[6] R. Jacobs, M. P. Polak, L. E. Schultz, H. Mahdavi, V. Honavar and D. Morgan, *arXiv*, 2024. <https://arxiv.org/abs/2409.06080>

[7] R. Vacareanu, V. A. Negru, V. Suciu and M. Surdeanu, in *First Conference on Language Modeling*, 2024. <https://openreview.net/forum?id=LzpaUxcNFK>

[8] A. K. Gupta and K. Raghavachari, *J. Chem. Theory Comput.*, 2022, **18**, 2132–2143. <https://doi.org/10.1021/acs.jctc.1c00504>

[9] S. Chithrananda, G. Grand and B. Ramsundar, *arXiv*, 2020. <https://arxiv.org/abs/2010.09885>

[10] J. Lu and Y. Zhang, *J. Chem. Inf. Model.*, 2022, **62**, 1376–1387. <https://doi.org/10.1021/acs.jcim.1c01467>

[11] D. Bhattacharya, H. J. Cassady, M. A. Hickner and W. F. Reinhart, *J. Chem. Inf. Model.*, 2024, **64**, 7086–7096. <https://doi.org/10.1021/acs.jcim.4c01396>

[12] G. Liu, M. Sun, W. Matusik, M. Jiang and J. Chen, *arXiv*, 2024. <https://arxiv.org/abs/2410.04223>

[13] S. Jia, C. Zhang and V. Fung, *arXiv*, 2024. <https://arxiv.org/abs/2406.13163>

[14] H. Jang, Y. Jang, J. Kim and S. Ahn, *arXiv*, 2024. <https://arxiv.org/abs/2410.03138>

[15] J. Lu, Z. Song, Q. Zhao, Y. Du, Y. Cao, H. Jia and C. Duan, *arXiv*, 2024. <https://arxiv.org/abs/2410.18136>

[16] A. Kristiadi *et al.*, in *Proceedings of the 41st International Conference on Machine Learning*, PMLR, 2024, vol. 235, pp. 25603–25622. <https://proceedings.mlr.press/v235/kristiadi24a.html>

[17] S. Miret and N. M. A. Krishnan, *arXiv*, 2024. <https://arxiv.org/abs/2402.05200>

[18] X. Ji, A. L. Nielsen and C. Heinis, *Angew. Chem. Int. Ed.*, 2023, **63**, 3. <https://doi.org/10.1002/anie.202308251>

[19] M. L. Merz *et al.*, *Nat. Chem. Biol.*, 2023, **20**, 624–633. <https://doi.org/10.1038/s41589-023-01496-y>

[20] P. Lewis *et al.*, *arXiv*, 2020. <https://arxiv.org/abs/2005.11401>

[21] A. Dubey *et al.*, *arXiv*, 2024. <https://arxiv.org/abs/2407.21783>

[22] M. Abdin *et al.*, *arXiv*, 2024. <https://arxiv.org/abs/2404.14219>

[23] A. M. Bran, S. Cox, O. Schilter, C. Baldassari, A. D. White and P. Schwaller, *arXiv*, 2023. <https://arxiv.org/abs/2304.05376>

[24] Y. Song *et al.*, *arXiv*, 2023. <https://arxiv.org/abs/2306.06624>

[25] H. Zhang, Y. Song, Z. Hou, S. Miret and B. Liu, *arXiv*, 2024. <https://arxiv.org/abs/2409.00135>

[26] D. A. Boiko, R. MacKnight, B. Kline and G. Gomes, *Nature*, 2023, **624**, 570–578. <https://doi.org/10.1038/s41586-023-06792-0>

[27] K. Darvish *et al.*, *arXiv*, 2024. <https://arxiv.org/abs/2401.06949>

[28] G. Tom *et al.*, *Chem. Rev.*, 2024, **124**, 9633–9732. <https://doi.org/10.1021/acs.chemrev.4c00055>

[29] H. Chase, *Langchain*, 2024. <https://github.com/langchain-ai/langchain>

[30] RDKit: Open-source cheminformatics, <http://www.rdkit.org>

[31] L. Yan *et al.*, *Br. J. Educ. Technol.*, 2023, **55**, 90–112. <https://doi.org/10.1111/bjet.13370>

[32] S. Wang *et al.*, *arXiv*, 2024. <https://arxiv.org/abs/2403.18105>

[33] E. Kasneci *et al.*, *Learn. Individ. Differ.*, 2023, **103**, 102274. <https://doi.org/10.1016/j.lindif.2023.102274>

[34] M. S. Schäfer, *J. Sci. Commun.*, 2023, **22**, 2. <https://doi.org/10.22323/2.22020402>

[35] M. Zaki, Jayadeva, Mausam and N. M. A. Krishnan, *arXiv*, 2023. <https://arxiv.org/abs/2308.09115>

[36] Anthropic, *The Claude 3 Model Family: Opus, Sonnet, Haiku*, 2024. <https://assets.anthropic.com/m/61e7d27f8c8f5919/original/Claude-3-Model-Card.pdf>

[37] A. Q. Jiang *et al.*, *arXiv*, 2024. <https://arxiv.org/abs/2401.04088>

[38] C. Draxl and M. Scheffler, *J. Phys. Mater.*, 2019, **2**, 036001. <https://doi.org/10.1088/2515-7639/ab13bb>

[39] A. Radford *et al.*, *arXiv*, 2022. <https://arxiv.org/abs/2212.04356>

[40] Y. Zhou, H. Liu, T. Srivastava, H. Mei and C. Tan, *arXiv*, 2024. <https://arxiv.org/abs/2404.04326>

[41] A. Abdel-Rehim *et al.*, *arXiv*, 2024. <https://arxiv.org/abs/2405.12258>

[42] S. Tong, K. Mao, Z. Huang, Y. Zhao and K. Peng, *Humanit. Soc. Sci. Commun.*, 2024, **11**, 1. <https://doi.org/10.1057/s41599-024-03407-5>

[43] I. Ciucă, Y.-S. Ting, S. Kruk and K. Iyer, *arXiv*, 2023. <https://arxiv.org/abs/2306.11648>

[44] Q. Liu *et al.*, *arXiv*, 2024. <https://arxiv.org/abs/2409.06756>

[45] O. Shir, *ChemRxiv*, 2024. <https://doi.org/10.26434/chemrxiv-2024-lf2xx>

[46] S. Yao, D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao and K. Narasimhan, in *Advances in Neural Information Processing Systems*, Curran Associates, Inc., 2023, vol. 36, pp. 11809–11822. [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/271db9922b8d1f4dd7aaef84ed5ac703-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/271db9922b8d1f4dd7aaef84ed5ac703-Paper-Conference.pdf)

[47] M. Shamsabadi, J. D’Souza and S. Auer, *arXiv*, 2024. <https://arxiv.org/abs/2401.10040>

[48] J. Dagdelen *et al.*, *Nat. Commun.*, 2024, **15**, 1. <https://doi.org/10.1038/s41467-024-45563-x>

[49] D. Xu *et al.*, *arXiv*, 2024. <https://arxiv.org/abs/2312.17617>

[50] J. Li, M. Zhang, N. Li, D. Weyns, Z. Jin and K. Tei, *ACM Trans. Adapt. Syst.*, 2024, **19**, 1–60. <https://doi.org/10.1145/3686803>

[51] Y. Ma *et al.*, *arXiv*, 2024. <https://arxiv.org/abs/2402.11451>

[52] OpenAI *et al.*, *arXiv*, 2023. <https://arxiv.org/abs/2303.08774>

[53] K. M. Jablonka *et al.*, *Digital Discovery*, 2023, **2**, 1233–1250. <https://doi.org/10.1039/d3dd00113j>

# Appendix: Individual Project Reports

## Table of Contents

1	Leveraging Orbital-Based Bonding Analysis Information in LLMs	15
2	Context-Enhanced Material Property Prediction (CEMPP)	17
3	MolFoundation: Benchmarking Chemistry LLMs on Predictive Tasks	19
4	3D Molecular Feature Vectors for Large Language Models	22
5	LLMSpectrometry	25
6	MC-Peptide: An Agentic Workflow for Data-Driven Design of Macrocyclic Peptides	27
7	Leveraging AI Agents for Designing Low Band Gap Metal-Organic Frameworks	30
8	How Low Can You Go? Leveraging SmallLLMs for Material Design	34
9	LangSim	37
10	LLMicroscopilot: assisting microscope operations through LLMs	40
11	T2DlIlla: Harnessing Language Model for Density Functional Theory (DFT) Parameter Suggestion	42
12	Materials Agent: An LLM-Based Agent with Tool-Calling Capabilities for Cheminformatics	45
13	LLM with Molecular Augmented Token	49
14	MaSTeA: Materials Science Teaching Assistant	52
15	LLMy-Way	55
16	WaterLLM: Creating a Custom ChatGPT for Water Purification Using Prompt-Engineering Techniques	57
17	yeLlowhaMMer: A Multi-modal Tool-calling Agent for Accelerated Research Data Management	60
18	LLMads	62
19	NOMAD Query Reporter: Automating Research Data Narratives	63
20	Speech-schema-filling: Creating Structured Data Directly from Speech	65
21	Leveraging LLMs for Bayesian Temporal Evaluation of Scientific Hypotheses	68
22	Multi-Agent Hypothesis Generation and Verification through Tree of Thoughts and Retrieval-Augmented Generation	71
23	ActiveScience	75
24	G-Peer-T: LLM Probabilities For Assessing Scientific Novelty and Nonsense	77
25	ChemQA: Evaluating Chemistry Reasoning Capabilities of Multi-Modal Foundation Models	78
26	LithiumMind - Leveraging Language Models for Understanding Battery Performance	81
27	KnowMat: Transforming Unstructured Material Science Literature into Structured Knowledge	84
28	Ontosynthesis	86
29	Knowledge Graph RAG for Polymer Simulation	89
30	Synthetic Data Generation and Insightful Machine Learning for High Entropy Alloy Hydrides	91
31	ChemSense: Are large language models aligned with human chemical preference?	93
32	GlossaGen	96

# 1 Leveraging Orbital-Based Bonding Analysis Information in LLMs

Authors: Katharina Ueltzen, Aakash Naik, Janine George

LLMs were recently demonstrated to perform well for materials property prediction, especially in the low-data limit [1, 2]. The Learning LOBSTERs team fine-tuned multiple Llama 3 models on textual descriptions of 1264 crystal structures to predict the highest-frequency peak in their phonon density of states (DOS) [3,4]. This target is relevant to the thermal properties of materials, and this target dataset is part of the MatBench benchmark project [3, 4].

The text descriptions were generated using two packages: the Robocrystallographer package [5] generates descriptions of structural features like bond lengths, coordination polyhedra, or structure type; it has recently emerged as a popular tool for materials property prediction models that require text input [6–8]. Further, text descriptions of orbital-based bonding analyses containing information on covalent bond strengths or antibonding states were generated using LobsterPy [9]. The data used here is available on Zenodo [10] and was generated as part of our previous study, in which the importance of such bonding information for the same target via an RF model was demonstrated [10].

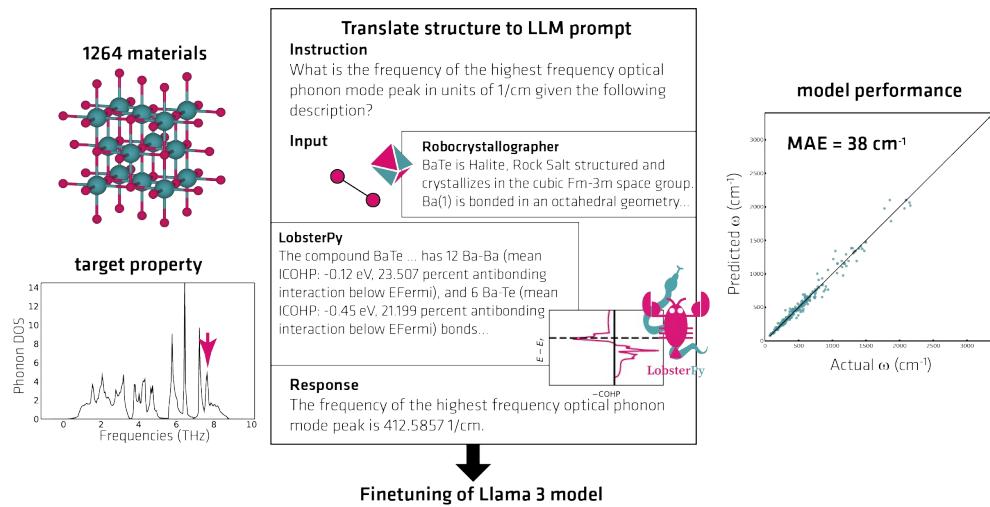


Figure 2: Schematic depicting the prompt for fine-tuning the LLM with Alpaca prompt format.

In the hackathon, one Llama model was fine-tuned with the Alpaca prompt format using both Robocrystallographer and LobsterPy text descriptions, and another one using solely Robocrystallographer input. Figure 2 depicts the prompt used to fine-tune an LLM to predict the last phonon DOS peak. The train/test-validation split was 0.64/0.2/0.16. The models were trained for 10 epochs with a validation step after each epoch. The textual output was converted back into numerical frequency values for the computation of MAEs and RMSEs. Our results show that including bonding-based information improved the model's prediction. The results also corroborate our previous finding that quantum-chemical bond strengths are relevant for this particular target property [10]. Both model performances (Robocrystallographer:  $44\text{ cm}^{-1}$ , Robocrystallographer+LobsterPy:  $38\text{ cm}^{-1}$ ) are comparable to other models of the MatBench test suite, with MAEs ranging from  $29\text{ cm}^{-1}$  to  $68\text{ cm}^{-1}$  at time of writing [11]. However, due to the time constraints of the hackathon, no five-fold cross-validation was implemented for our model.

Although the preliminary results seem very promising, the models have not yet been exhaustively analyzed or improved. As the prediction of a numerical value and not its text embedding is of interest to our task, further model adaptation might be beneficial. For example, Rubungo et al. [7] modified T5 [12], an encoder-decoder model, for regression tasks by removing its decoder and adding a linear layer on top of its encoder. Halving the number of model parameters allowed them to fine-tune on longer input sequences, improving model performance [7].

Easy-to-use packages like Unsloth [13] allowed us to integrate our materials data into fine-tuning an LLM for property prediction with very limited resources and time.

## References

- [1] K. M. Jablonka, P. Schwaller, A. Ortega-Guerrero, B. Smit, *Nat Mach Intell*, 2024, 6, 161–169.
- [2] K. Choudhary, *J. Phys. Chem. Lett.*, 2024, 6909–6917.
- [3] G. Petretto, S. Dwarakanath, H. P.C. Miranda, D. Winston, M. Giantomassi, M. J. van Setten, X. Gonze, K. A. Persson, G. Hautier, G.-M. Rignanese, *Sci Data*, 2018, 5, 180065.
- [4] A. Dunn, Q. Wang, A. Ganose, D. Dopp, A. Jain, *npj Comput Mater*, 2020, 6, 1–10.
- [5] A. M. Ganose, A. Jain, *MRS Communications*, 2019, 9, 874–881.
- [6] H. M. Sayeed, S. G. Baird, T. D. Sparks, 2023, DOI 10.26434/chemrxiv-2023-3q8wj.
- [7] A. N. Rubungo, C. Arnold, B. P. Rand, A. B. Dieng, 2023, DOI 10.48550/arXiv.2310.14029.
- [8] V. Moro, C. Loh, R. Dangovski, A. Ghorashi, A. Ma, Z. Chen, S. Kim, P. Y. Lu, T. Christensen, M. Soljačić, 2024, DOI 10.48550/arXiv.2312.00111.
- [9] A. A. Naik, K. Ueltzen, C. Ertural, A. J. Jackson, J. George, *Journal of Open Source Software*, 2024, 9, 6286.
- [10] A. A. Naik, C. Ertural, N. Dhamrait, P. Benner, J. George, 2023, DOI 10.5281/zenodo.8091844.
- [11] The Matbench Test Suite, Phonon dataset as per 12.07.2024 <https://matbench.materialsproject.org/Leaderboards>
- [12] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, *Journal of Machine Learning Research*, 2020, 21, 1–67.
- [13] The Unsloth package, <https://github.com/unslothai/uns> 2024.

## 2 Context-Enhanced Material Property Prediction (CEMPP)

Authors: Federico Ottomano, Elena Patyukova, Judith Clymo, Dmytro Antypov, Chi Zhang, Aritra Roy, Piyush Ranjan Maharana, Weijie Zhang, Xuefeng Liu, Erik Bitzek

### 2.1 Introduction

The Liverpool Materials team sought to improve composition-based property prediction models for novel materials by providing natural language input alongside the chemical composition of interest. In doing so, we leverage the ability of large language models (LLMs) to capture the broader context relevant to the composition. We demonstrate that enriching materials representation can be beneficial where training data is limited. Code to reproduce the experiments described below is available at [https://github.com/fedeotto/lpoolmat\\_llms](https://github.com/fedeotto/lpoolmat_llms)

Our experiments are based on Roost [1], a deep neural network for predicting inorganic material properties from their composition. Roost consists of a graph attention network that creates an embedding of the input composition (which we will enrich with context embedding), and a residual network that acts on this embedding to predict the target property value (Figure 3, top).

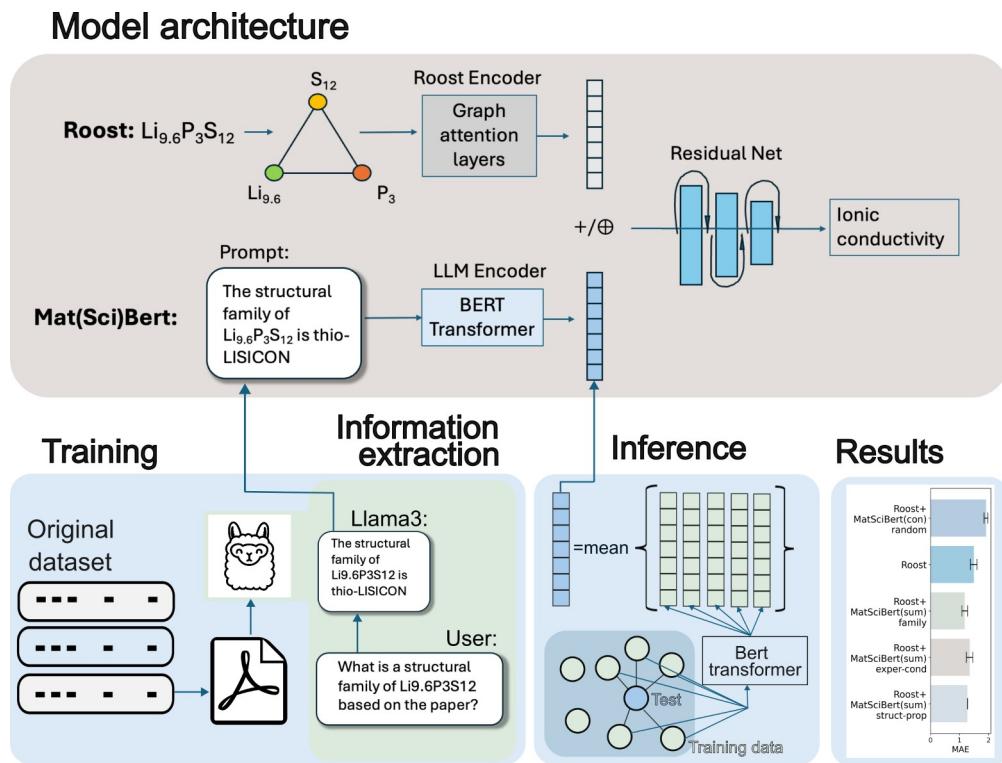


Figure 3: Model architecture and the schema of the second experiment. Material composition is encoded with Roost encoder, additional information extracted from cited paper with Llama3, and encoded with Mat(Sci)Bert. Composition and LLM embeddings are aggregated and passed through the Residual Net projection head to predict the property. At the inference stage, the average LLM embedding from 5 nearest neighbors in composition space is taken. Results show MAE for adding different types of context (top to bottom): adding random context; not adding context; adding consistently structured context for chemical and structural family (data extracted by humans); adding automatically extracted context for experimental conditions and structure-property relationship.

## 2.2 Experiment 1:Using latent knowledge

We prompt two LLMs trained on materials science literature, MatBert [2] and MatSciBert [3], to directly attempt the prediction task ("What is the property of material?"). The embedding of the response and Roost's composition embedding are aggregated (via summation or concatenation) and passed to the residual network.

We consider two datasets: matbench-perovskites [4] has a target property of formation energy and Li-ion conductors [5] dataset has a target property of ionic conductivity. The former has low stoichiometric diversity in the inputs (all entries have a general formula of ABX3) and the latter is limited in size (only 403 distinct compositions), making the prediction tasks especially challenging. We observe a 26-32% decrease in mean absolute error (MAE) for all four settings tested (two LLMs and two aggregation schemes) in the Li-ion conductors task, and a 2-4% decrease in MAE in the matbench-perovskites task.

## 2.3 Experiment 2:Using knowledge from literature

We obtain information from the papers reporting materials in the Li-ion conductors dataset by using PyMuPDF [6] parser and scanning for keywords to find relevant sections. We prompt Llama3-8b-instruct [7] to extract and summarize specific information from the text, creating an embedding from the response (Figure 3).

Given our goal of predicting properties for compositions not yet synthesized, and therefore not referenced in any academic text, for testing and inference we average the summary embeddings of the five closest materials in the training set to the composition of interest. We define the distance between compositions using the Element Movers Distance [8].

We find that providing context embeddings generated from automatically extracted experimental conditions and structure-property relationships reduce MAE by 3–10% and 10–15% respectively. Using context embeddings based on information extracted by human experts [9] from the same academic papers reduces MAE by 17–21%. In contrast, using random embeddings as context increases MAE by 7–28%. We conclude that automatically extracted context can aid property prediction for new compositions. The performance boost is less than that achieved using human-extracted and consistently structured context, which highlights both the further potential of our method and the harmful effect of noise in the automatically extracted text.

## References

- [1] R. Goodall, A. J. et al., "Predicting materials properties without crystal structure: deep representation learning from stoichiometry", *Nature Communications*, vol. 11, no. 6280, 2020.
- [2] A. Trewartha, et al., 'Quantifying the advantage of domain-specific pre-training on named entity recognition tasks in materials science', *Patterns*, vol. 3, no. 8, 2022.
- [3] V. Gupta, et al., *MatSciBERT: A materials domain language model for text mining and information extraction*, *npj Computational Materials*, vol. 8, no. 1, 2022.
- [4] Matbench Perovskites dataset provided by the Materials Project [https://ml.materialsproject.org/projects/matbench\\_perovskites.json.gz](https://ml.materialsproject.org/projects/matbench_perovskites.json.gz)
- [5] Li-ion Conductors Database, <https://pcwww.liv.ac.uk/~sd30/lmds/LiIonDatabase.html>
- [6] PyMuPDF, <https://pypi.org/project/PyMuPDF>
- [7] AI@Meta, *Llama 3 Model Card*, 2024, [https://github.com/meta-llama/llama3/blob/main/MODEL\\_CARD.md](https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md)
- [8] Hargreaves et al., *The Earth Mover's Distance as a Metric for the Space of Inorganic Compositions*, *Chem. Mater.* 2020.
- [9] Hargreaves et al., *A Database of Experimentally Measured Lithium Solid Electrolyte Conductivities Evaluated with Machine Learning*, *npj Computational Materials* 2022.

### 3 MolFoundation: Benchmarking Chemistry LLMs on Predictive Tasks

Authors: Hassan Harb, Xuefeng Liu, Anastasiia Tsybala, Oleksandr Narykov, Dana O'Connor, Shagun Maheshwari, Stanley Lo, Archit Vasan, Zartashia Afzal, Kevin Shen

#### 3.1 Summary

The MolFoundation team is focused on enhancing the prediction capabilities of pre-trained large language models (LLMs) such as ChemBERTa and T5-Chem for specific molecular properties, utilizing the QM9 database. Targeting properties like dipole moment and zero-point vibrational energy (ZPVE), our approach involves retraining the last layer of these models with a selected dataset and fine-tuning the embeddings to refine accuracy. After making predictions and evaluating performance, our results indicate that T5-Chem outperforms ChemBERTa. Additionally, we found little difference between finetuned and pre-trained results, suggesting that the computationally expensive task of finetuning may be avoided.

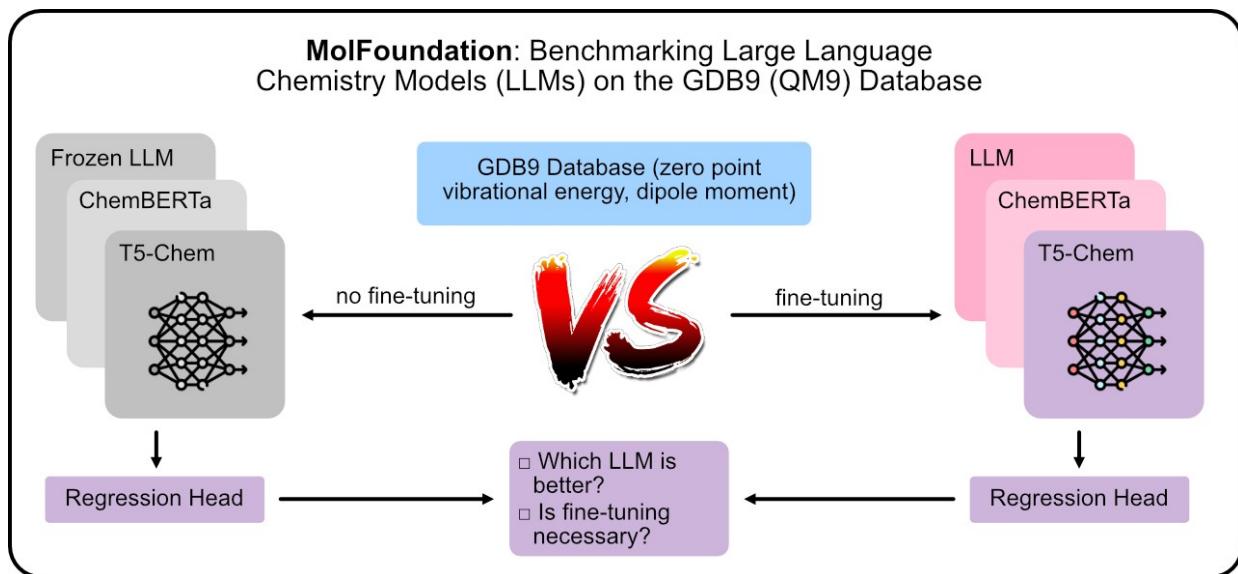


Figure 4

#### 3.2 Methods

The models are downloaded from HuggingFace (ChemBERT and T5-Chem). Using the provided code we can tokenize our datasets (QM9). The datasets must contain SMILES and we checked that the tokenizer had all the necessary tokens for our datasets. To make the LLMs compatible with the regression tasks in the QM9 dataset, we froze the LLM embeddings and fine-tuned on the regression layer. Training on the full LLM end-to-end is infeasible given our resources, so training on a single linear layer was much more efficient.

#### 3.3 Results

We compare the out-of-the-box (pre-trained) and fine-tuned ChemBERT and T5-Chem to predict all the molecules' zero-point vibrational energy (ZPVE) in the QM9 dataset. We hypothesized that the fine-tuned models would perform better. However, across all models, there is no significant improvement in the fine-tuned models as measured by  $R^2$  (Figure 6, Figure 7). We noticed that the LLMs required approximately 100K datapoints to show improvements in the modeling performance, indicating a saturation regime for the models. Lastly, the T5-Chem model performs significantly better than ChemBERT.

```

# Load model directly
from transformers import AutoTokenizer, AutoModelForMaskedLM

tokenizer = AutoTokenizer.from_pretrained("seyonec/ChemBERTa-zinc-base-v1")
model = AutoModelForMaskedLM.from_pretrained("seyonec/ChemBERTa-zinc-base-v1")

```

Figure 5

## T5-Chem Pretrained vs T5-Chem Finetuned Results (ZPVE)

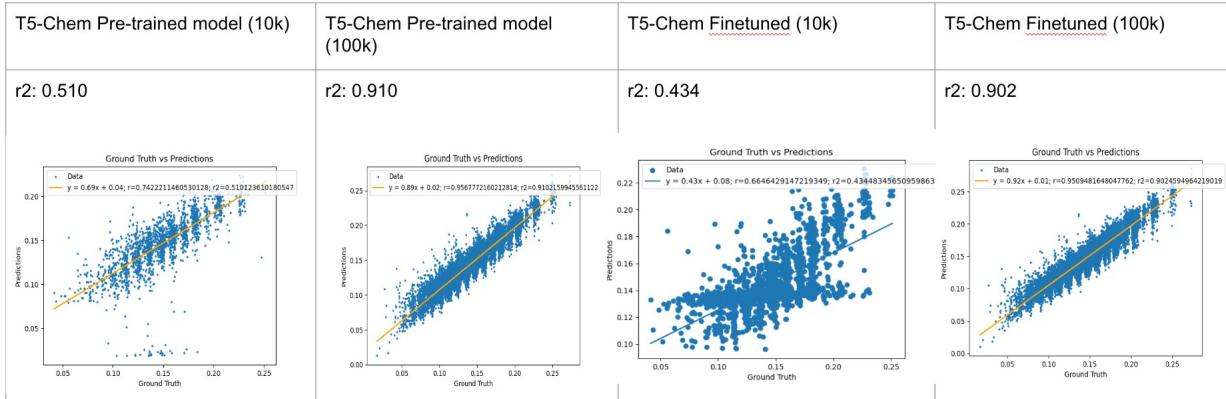


Figure 6: Model comparison of the pre-trained and fine-tuned T5-Chem on zero-point vibrational energy (ZPVE).

## ChemBERTa Pretrained vs ChemBERTa Finetuned Results (ZPVE)

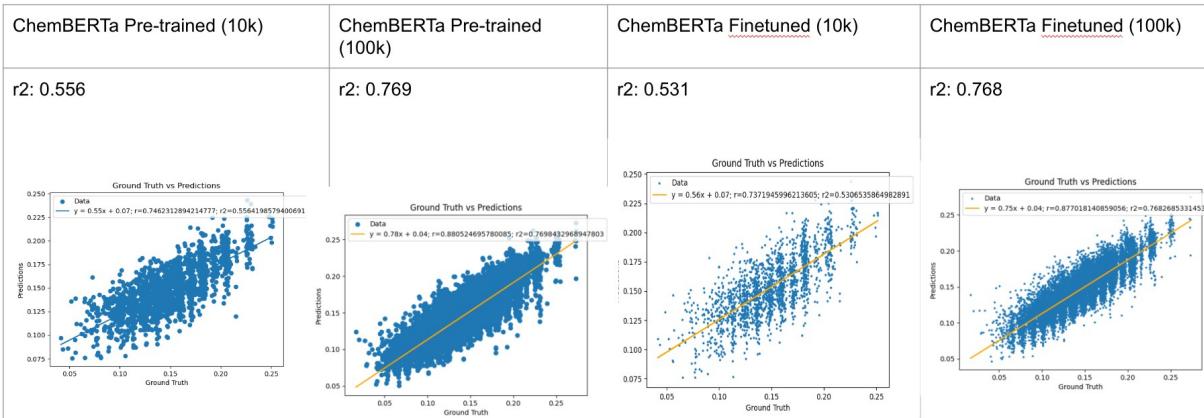


Figure 7: Model comparison of the pre-trained and fine-tuned ChemBERTa on zero-point vibrational energy (ZPVE).

### 3.4 Transferability + Outlook

From our experiments, it was unexpected to find that there was a lack of transfer learning from the pre-trained LLMs to the predictive tasks on the QM9 dataset. Nevertheless, we need more experimentation on

different datasets, models, and fine-tuning strategies (i.e., end-to-end retraining, multi-task fine-tuning) to determine the efficacy of LLMs on chemistry prediction tasks.

MolFoundation Github: <https://github.com/shagunm1210/MolFoundation>

## References

[1] Kristiadi, A.; Strieth-Kalthoff, F.; Skreta, M.; Poupart, P.; Aspuru-Guzik, A.; Pleiss, G. A Sober Look at LLMs for Material Discovery: Are They Actually Good for Bayesian Optimization Over Molecules? arXiv May 28, 2024. <http://arxiv.org/abs/2402.05015> (accessed 2024-06-29).

## 4 3D Molecular Feature Vectors for Large Language Models

Authors: Jan Weinreich, Ankur K. Gupta, Amirhossein D. Naghdi, Wibe A. de Jong, Alishba Imran

Link to code repo: <https://github.com/janweinreich/geometric-genuses>

Direct link to tutorial: <https://github.com/janweinreich/geometric-genuses/blob/main/tutorial.ipynb>

Accurate chemical property prediction is a central goal in computational chemistry and materials science. While quantum chemistry methods offer high precision, they often suffer from computational bottlenecks. Large language models (LLMs) have shown promise as a computationally efficient alternative [1]. However, common string-based molecular representations like SMILES and SELFIES, despite their success in LLM applications, inherently lack 3D geometric information. This limitation hinders their applicability in predicting properties for different conformations of the same molecule - a capability essential for practical applications such as crystal structure prediction and molecular dynamics simulations. Moreover, recent studies have demonstrated that naive encoding of geometric information as numerical data can negatively impact LLM prediction accuracy [2].

Molecular and materials property prediction typically leverages engineered feature vectors, such as those utilized in quantitative structure-activity relationship (QSAR) models. In contrast, physics-based representations, which center on molecular geometry due to their direct relevance to the Schrödinger equation, have demonstrated efficacy in various deep learning architectures [3–5]. This research investigates new strategies for encoding 3D molecular geometry for LLMs. We hypothesize that augmenting the simple SMILES representation with geometric features could enable the integration of complementary information from both modalities, ultimately enhancing the predictive power of LLMs in the context of molecular properties.

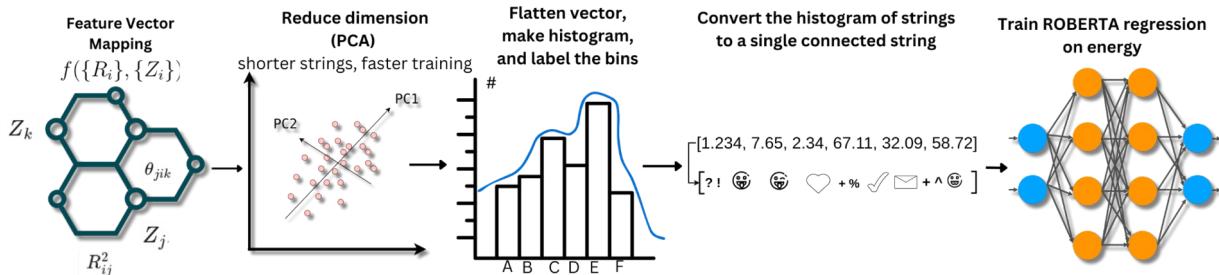


Figure 8: Schematic representation of the training process for a regression model illustrating the novel string-based encoding of 3D molecular geometry for LLM-based energy prediction. The workflow involves (1) Computation of high-dimensional feature vectors representing the 3D molecular geometry of each conformer. (2) Dimensionality reduction to obtain a more compact representation. (3) Conversion of the reduced vectors into a histogram, where unique string characters are assigned to each bin. (4) Input of the resulting set of strings (one per conformer) into an LLM for energy prediction.

To achieve the integration of geometric information, we begin by computing a high-dimensional feature vector for each molecular geometry. While any representation capable of faithfully encoding the 3D structure of a molecule could, in principle, be utilized, this work specifically employs a previously established geometry-dependent representation [6] based on many-body expansions of distances and angles. As a benchmark dataset, we utilize a diverse set of ethanol and benzene structures generated using molecular dynamics calculations utilizing density functional theory (DFT) [7, 8]. The target property we aim to predict is the absolute energy of each conformer. The energy scale is shifted such that the lowest energy conformer within the dataset is assigned an energy value of zero. We transform our high-dimensional geometric feature vectors into string representations suitable for LLM training. First, we apply principal component analysis (PCA) to reduce the dimensionality of the vectors. This step is crucial for generating compact string representations that can be efficiently processed by LLMs. Next, we compute histograms of the reduced vectors, ensuring consistent binning intervals across all dimensions. Each bin is uniquely labeled with a distinct character, and negative values are prefixed with a special character, following the approach of [9]. Finally, we count

the occurrences of values within each bin, effectively converting the numerical vectors into strings. These string representations serve as the input for training a RoBERTa regression model.

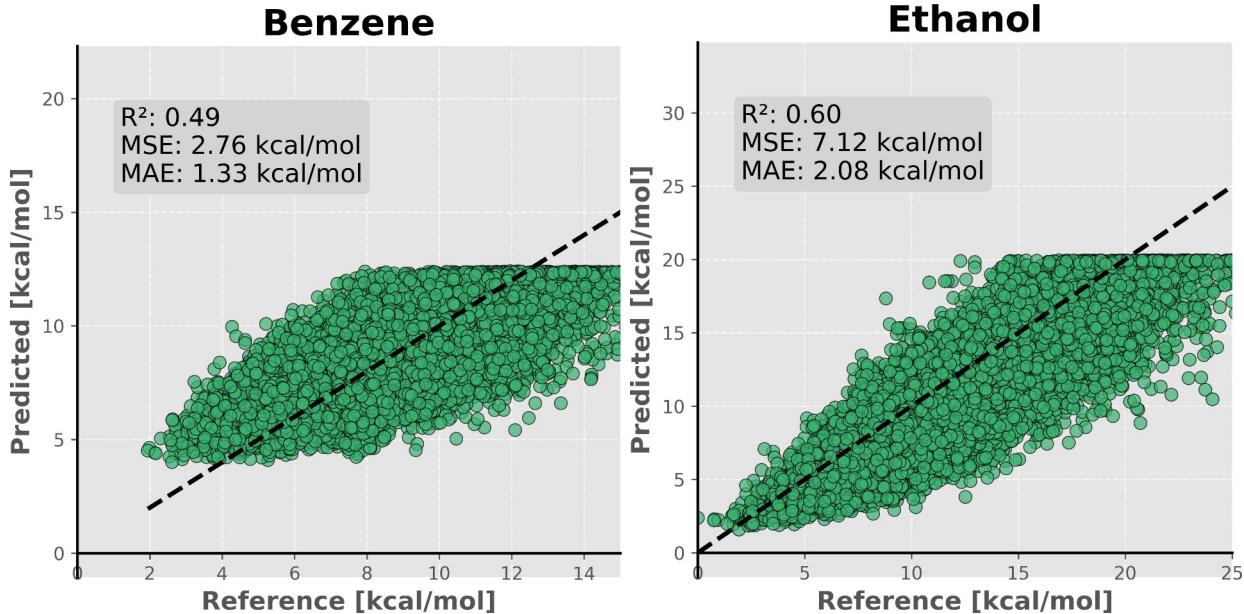


Figure 9: Performance of an LLM in predicting total energies of benzene and ethanol structures, where the model was trained on a large dataset of MD-generated configurations. Each point represents a different molecular structure sampled from MD simulations [7].

The final step involves randomly partitioning the dataset into 80% training and 20% testing sets. The string-based representations are then employed to train a RoBERTa model augmented with a regression head. In Figure 9, we showcase scatter plots illustrating the predicted versus true total energies for benzene (trained for 4 epochs) and ethanol (trained for 20 epochs) using a dataset of 80,000 molecular configurations for each molecule. Our results do not yet attain the accuracy levels of state-of-the-art equivariant neural networks like MACE, which reports a mean absolute error (MAE) of 0.009 kcal/mol for benzene [10]. Nonetheless, it is important to underscore that this represents a novel capability for LLMs, which were previously unable to process and predict properties of 3D molecular structures only

differing by bond rotations. This initial investigation paves the way for advancements through the exploration of alternative string encoding schemes of numerical vectors in combination with larger LLMs.

## Acknowledgements

A.K.G. and W.A.D. acknowledge funding for this project from the U.S. Department of Energy (DOE), Office of Science, Office of Basic Energy Sciences, through the Rare Earth Project in the Separations Program at Lawrence Berkeley National Laboratory under Contract DE-AC02-05CH11231.

J.W. thanks EPFL for computational resources and NCCR Catalysis (grant number 180544), a National Centre of Competence in Research funded by the Swiss National Science Foundation for funding as well as the Laboratory for Computational Molecular Design.

## References

[1] Jablonka, K. M., Ai, Q., Al-Feghali, A., Badhwar, S., Bocarsly, J. D., Bran, A. M., ... & Blaiszik, B. (2023). 14 examples of how LLMs can transform materials science and chemistry reflection on a large language model hackathon. *Digital Discovery*, 2(5), 1233-1250.

- [2] Alampara, N., Miret, S., & Jablonka, K. M. (2024). MatText: Do Language Models Need More than Text & Scale for Materials Modeling? arXiv preprint [arXiv:2406.17295](https://arxiv.org/abs/2406.17295).
- [3] Batzner, S., Musaelian, A., Sun, L., Geiger, M., Mailoa, J. P., Kornbluth, M., ... & Kozinsky, B. (2022). E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13(1), 2453.
- [4] Gupta, A. K., & Raghavachari, K. (2022). Three-dimensional convolutional neural networks utilizing molecular topological features for accurate atomization energy predictions. *Journal of Chemical Theory and Computation*, 18(4), 2132-2143.
- [5] In-Context Learning of Physical Properties: Few-Shot Adaptation to Out-of-Distribution Molecular Graphs. Grzegorz Kaszuba, Amirhossein D. Naghdi, Dario Massa, Stefanos Papanikolaou, Andrzej Jaszkiewicz, Piotr Sankowski, <https://arxiv.org/abs/2406.01808>
- [6] Khan, D., Heinen, S., & von Lilienfeld, O. A. (2023). Kernel based quantum machine learning at record rate: Many-body distribution functionals as compact representations. *Journal of Chemical Physics*, 159(034106).
- [7] Chmiela, S., Vassilev-Galindo, V., Unke, O. T., Kabylda, A., Sauceda, H. E., Tkatchenko, A., & Müller, K. R. (2023). Accurate global machine learning force fields for molecules with hundreds of atoms. *Science Advances*, 9(2), <https://doi.org/10.1126/sciadv.adf0873>
- [8] Bowman, J. M., Qu, C., Conte, R., Nandi, A., Houston, P. L., & Yu, Q. (2022). The MD17 datasets from the perspective of datasets for gas-phase “small” molecule potentials. *The Journal of chemical physics*, 156(24).
- [9] Weinreich, J., & Probst, D. (2023). Parameter-Free Molecular Classification and Regression with Gzip. *ChemRxiv*.
- [10] Batatia, I., Kovacs, D. P., Simm, G., Ortner, C., & Csányi, G. (2022). MACE: Higher order equivariant message passing neural networks for fast and accurate force fields. *Advances in Neural Information Processing Systems*, 35, 11423-11436.

## 5 LLMSpectrometry

Authors: *Tyler Josephson, Fariha Agbere, Kevin Ishimwe, Colin Jones, Charishma Puli, Samiha Sharlin, Hao Liu*

### 5.1 Introduction

Nuclear Magnetic Resonance (NMR) spectroscopy is a chemical characterization technique that uses oscillating magnetic fields to characterize the structure of molecules. Different atoms in molecules resonate at different frequencies based on their local chemical environments. The resulting NMR spectrum can be used to infer interactions between particular atoms in a molecule and determine the molecule's entire structure. Solving NMR spectral tasks is critical, as multiple aspects, such as the number, intensity, and shape of signals, as well as chemical shifts, need to be considered.

Machine learning tools, including the Molecular Transformer [1], have been used to learn a function to map spectrum to structure, but these require thousands to millions of labeled examples [2], far more than what humans typically encounter when learning to assign spectra. In contrast, we recognize NMR spectral tasks as being fundamentally about multi-step reasoning, and we aim to explore the reasoning capabilities of large language models (LLMs) for solving this task.

The project aims to investigate the capabilities of Large Language Models (LLMs), specifically GPT-4, for NMR structure determination. In particular, we were interested in evaluating whether GPT-4 could use chain-of-thought reasoning [3] with a scratchpad to evaluate the components of the spectra and synthesize an answer in the form of a molecule. Interpreting NMR data is crucial for organic chemists; an AI-assisted tool could benefit the pharmaceutical or food industry, as well as forensic science, medicine, research, and teaching.

### 5.2 Method

We manually gathered 19 experimental  $^1\text{H}$  NMR data from the Spectral Database for Organic Compounds (SDBS) website [4]. We then combined Python scripts, LangChain, and API calls to GPT-4 to automate structure elucidation. The components of the model are shown in Figure 10. First, NMR peak data and the chemical formula were formatted as text and inserted into a prompt. This prompt instructs the LLM to reason about the data step-by-step while using a scratchpad to record its "thoughts," then report its answer according to an output template. We then used Python to parse the output and compare the answer to the true answer by matching with chemical names from the NIST Chemistry WebBook.

### 5.3 Results

Our 2024 LLM Hackathon for Applications in Materials and Chemistry (May 2024) results found GPT-4 successful in just 3 of the 19 NMR datasets. The scratchpad reveals how GPT-4 follows the prompt and systematically approaches the problem. It analyzes the molecular formula and carefully reads peak values and intensity to predict the molecule. The model correctly identified nonanal, ethanol, and acetic acid - 3 relatively simple molecules with few structural isomers. Incorrect answers included more complex molecules, with significant branching, many functional groups, and aromatic rings, leading to more structural isomers consistent with the chemical formula.

### 5.4 Conclusions and Outlook

NMR spectral tasks challenge students to develop complex problem solving skills, and these also prove to be difficult in a zero-shot chain-of-thought prompting setting for GPT-4, with only 3/19 spectra solved correctly. We noticed interesting patterns of (apparent) reasoning, and we speculate significant performance improvements are possible. Prompting strategies can be explored more systematically (including few-shot and many-shot approaches), as well as embedding the LLM's answers inside an iterative loop, so it can self-correct simple mistakes such as generating a molecule with an incorrect chemical formula.

Since the hackathon, we identified a better benchmark dataset: NMR-Challenge.com [3], an interactive website with exercises categorized as Easy, Medium, and Hard. They also have data on human performance,

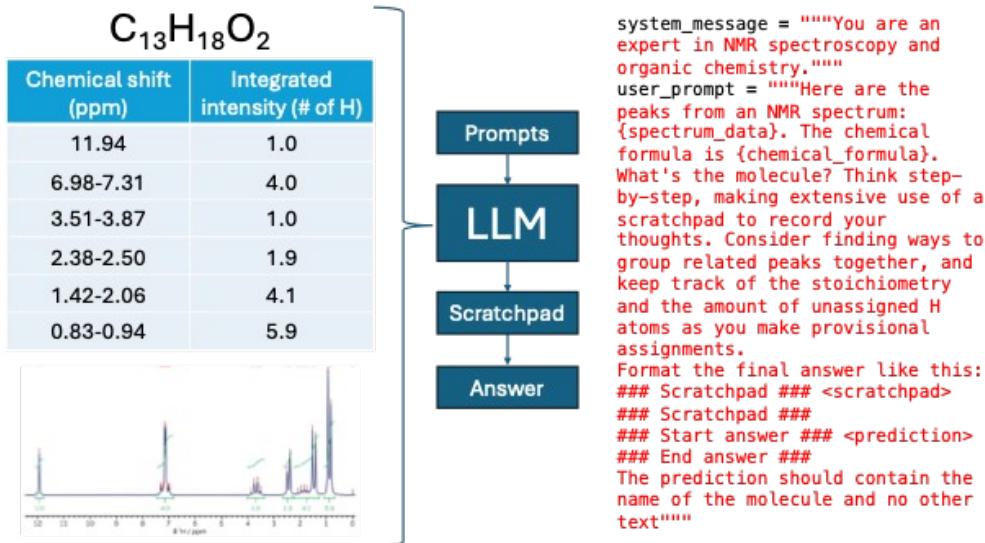


Figure 10: Scheme of the systemData is first converted into text format, with peak positions and intensities represented as (x,y) pairs. These are passed into an LLM prompt, which is tasked to use a scratchpad as it reasons about the data and the formula, before providing a final answer.

which can enable comparison of GPT-4 to humans. Further analysis of proximity of incorrect answers to correct answers would provide more granular information about the performance of the AI, for example, an aromatic molecule with correct substituents in incorrect locations is further from the right answer than a molecule with incorrect substituents. We think this could form a useful addition to existing LLM benchmarks, for evaluating chemistry knowledge intertwined with complex multistep reasoning.

Datasets, code, and results are available at: <https://github.com/ATOMSLab/LLMSpectroscopy>

## References

- [1] Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C., Bekas, C., Lee, A. A., *ACS Central Sci.*, 2019, Vol. 5, No. 9, 1572-1583. <https://pubs.acs.org/doi/10.1021/acscentsci.9b00576>
- [2] Alberts, M., Zipoli, F., Vaucher, A. C., *ChemRxiv Preprint*. <https://doi.org/10.26434/chemrxiv-2023-8wxcz>
- [3] Wei, J., Wang X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. *NeurIPS 2022*. <https://arxiv.org/abs/2201.11903>
- [4] Yamaji, T., Saito, T., Hayamizu, K., Yanagisawa, M., Yamamoto, O., Wasada, N., Someno, K., Kinugasa, S., Tanabe, K., Tamura, T., Hiraishi, J., 2024. <https://sdbs.db.aist.go.jp>
- [5] Socha, O., Osifova, Z., Dracinsky, M., *J. Chem. Educ.*, 2023, Vol. 100, No. 2, 962-968. <https://pubs.acs.org/doi/10.1021/acs.jchemed.2c01067>
- [6] <https://github.com/ATOMSLab/LLMSpectroscopy>

## 6 MC-Peptide: An Agentic Workflow for Data-Driven Design of Macro cyclic Peptides

Authors: *Andres M. Bran, Anna Borisova, Marcel M. Calderon, Mark Tropin, Rob Mills, Philippe Schwaller*

### 6.1 Introduction

Macro cyclic peptides (MCPs) are a class of compounds composed of cyclized chains of amino acids forming a macro cyclic structure. They are promising for having improved binding affinity, specificity, proteolytic stability, and enhanced membrane permeability compared to linear peptides [1]. Their unique properties make them highly suitable for drug development, enabling the creation of therapeutics that can address currently unmet medical needs [1]. Indeed, it has been shown that MCPs of up to 12 amino acids show great promise for permeating cellular membranes [2]. Despite the more constrained chemical space offered by this class of molecules, their design remains a challenging issue due to the vast space of amino acid combinations.

One important parameter of MCPs is permeability, which determines how well the structure can permeate into cells, making it a relevant factor in assessing the MCP's pharmacological activity. However, data for MCPs and their permeabilities is scattered across scientific papers that report data without consensus on reporting form, making it challenging for systems to compile and use this raw data.

Here, we introduce MC-Peptide, an LLM-based agentic workflow created for tackling this issue in an end-to-end fashion. As shown in Figure 1, MC-Peptide's main goal is to produce suggestions of novel MCPs, following from a reasoning process involving: (i) understanding of the design objective, (ii) gathering of relevant scientific information, (iii) data extraction from papers, and (iv) inference based on the extracted information. MC-Peptide leverages advances in LLMs such as semantic search grammar-constrained generation, and agentic architectures, ultimately yielding candidate MCP variants with enhanced predicted permeability in comparison to reference designs.

### 6.2 Methods

We implemented the basic building blocks for the pipeline described in Figure 11, with the most important components being document retrieval, structured data extraction, and in-context learning for design.

**Document Retrieval** An important part of this pipeline is the retrieval of relevant documents. As shown in Figure 11, the pipeline developed here does not rely on a stand-alone predictive system, but rather aims to leverage existing knowledge in the scientific literature, in an on-demand fashion. This is known as retrieval-augmented generation (RAG) [8], and one of its key components is a relevance-based retrieval system. By integrating the Semantic Scholar API, a service that provides access to AI-based search of scientific documents, the pipeline is able to create and refine a structured knowledge base from papers.

**Structured Data Extraction** To employ the unstructured data from the previous step, a retrieval-augmented system was designed that retrieves sections of papers and uses them as context for the generation of structured data objects with predefined grammars, which can be used to constrain the decoding of LLMs [3], ensuring that the resulting output follows the same structure and format. This technique also mitigates hallucination, as it prevents the LLM from generating unnecessary and potentially misleading information [9].

**In-Context Learning** The extracted data is then leveraged through the in-context learning capabilities of LLMs [4], allowing the models to learn from few data points given as context in the prompt and generate output based on that. This capability has been extensively explored elsewhere [5]. Here we show that, for the specific task of modifying MCPs to improve permeability, LLMs perform well, as assessed by a surrogate random forest model.

### 6.3 Conclusions

We present MC-Peptide, a novel agentic workflow for designing macrocyclic peptides. The system is built from a few main components: data collection, peptide extraction, and peptide generation. We evaluate the peptide permeability of newly generated peptides with respect to the initial structures found in reference articles.

The resulting system shows that LLMs can be successfully leveraged for automating multiple aspects of peptide design, yielding an end-to-end generative tool that researchers can utilize to accelerate and enhance their experimental workflows. Furthermore, this workflow can be extended by adding more specialized modules (e.g., for increasing the diversity of information sources). The modular design of MC-Peptide ensures extendability to more design objectives and input sources, as well as other domains where data is reported in an unstructured fashion in papers, such as materials science and organic chemistry.

The code for this project has been made available at: <https://github.com/doncamilom/mc-peptide>

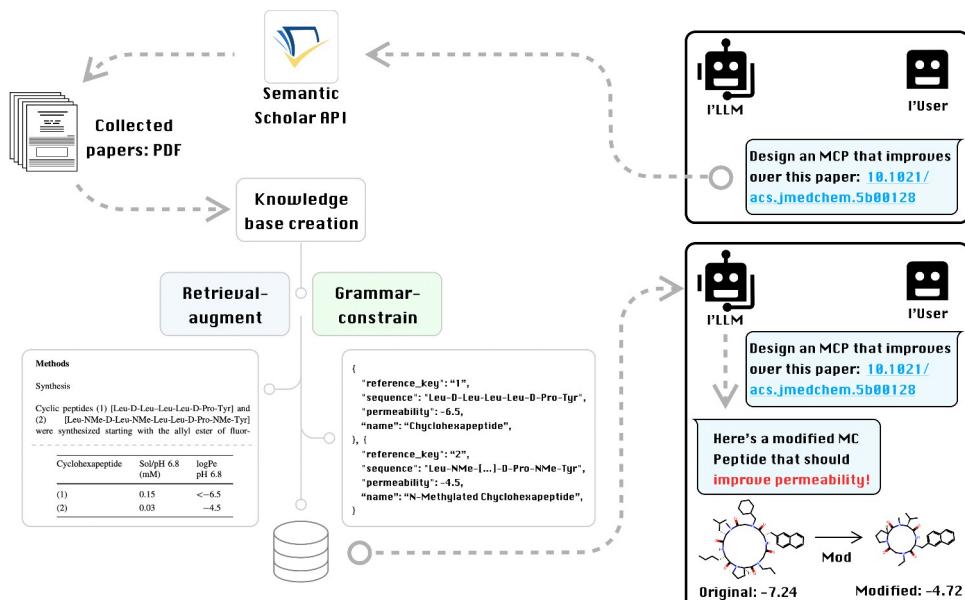


Figure 11: MC-Peptide: Pipeline implemented in this work. The example illustrates a user request, followed by retrieval from the Semantic Scholar API, and the creation of a knowledge base. MCPs and permeabilities are extracted from [7]. The pipeline finishes with an LLM proposing modifications to an MCP, increasing its permeability.

## References

- [1] Ji, X., Nielsen, A. L., Heinis, C., "Cyclic peptides for drug development," *Angewandte Chemie International Edition*, 2024, 63(3), e202308251.
- [2] Merz, M.L., Habeshian, S., Li, B. et al., "De novo development of small cyclic peptides that are orally bioavailable," *Nat Chem Biol*, 2024, 20, 624–633. <https://doi.org/10.1038/s41589-023-01496-y>
- [3] Beurer-Kellner, L., et al., "Guiding LLMs The Right Way: Fast, Non-Invasive Constrained Generation," *ArXiv*, 2024, abs/2403.06988.
- [4] Dong, Q., et al., "A survey on in-context learning," *ArXiv*, 2022, [arXiv:2301.00234](https://arxiv.org/abs/2301.00234).
- [5] Agarwal, R., et al., "Many-shot in-context learning," *ArXiv*, 2024, [arXiv:2404.11018](https://arxiv.org/abs/2404.11018).

[6] Kristiadi, A., et al., "A Sober Look at LLMs for Material Discovery: Are They Actually Good for Bayesian Optimization Over Molecules?," *ArXiv*, 2024, [arXiv:2402.05015](https://arxiv.org/abs/2402.05015).

[7] Lewis, I., Schaefer, M., Wagner, T. et al., "A Detailed Investigation on Conformation, Permeability and PK Properties of Two Related Cyclohexapeptides," *Int J Pept Res Ther*, 2015, 21, 205–221. <https://doi.org/10.1007/s10989-014-9447-3>

[8] Lewis, P., et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," *ArXiv*, 2020, [abs/2005.11401](https://arxiv.org/abs/2005.11401).

[9] Béchard, P., Ayala, O. M., "Reducing hallucination in structured outputs via Retrieval-Augmented Generation," *ArXiv*, 2024, [abs/2404.08189](https://arxiv.org/abs/2404.08189).

## 7 Leveraging AI Agents for Designing Low Band Gap Metal-Organic Frameworks

Authors: Sartaaj Khan, Mahyar Rajabi, Amro Aswad, Seyed Mohamad Moosavi, Mehrad Ansari

### 7.1 Introduction

Metal-organic frameworks (MOFs) are known to be excellent candidates for electrocatalysis due to their large surface area, high adsorption capacity at low CO<sub>2</sub> concentrations, and the ability to fine-tune the spatial arrangement of active sites within their crystalline structure [1]. Low band gap MOFs are crucial as they efficiently absorb visible light and exhibit higher electrical conductivity, making them suitable for photocatalysis, solar energy conversion, sensors, and optoelectronics. In this work, we aim at using chemistry-informed ReAct [2] AI Agents to optimize the band gap property of MOFs. The overview of the workflow is presented in Figure 12a. The agent inputs a textual representation of the initial MOF structure as a SMILES (Simplified Molecular Input Line-Entry System) string representation, and a short description of the property optimization task (i.e., reducing band gap), all in natural language text. This is followed by an iterative closed-loop suggestion of new MOF candidates with a lower band gap with uncertainty assessment, by making adjustments to the initial MOF given a set of design guidelines automatically obtained from the scientific literature. Detailed analysis of this methodology applied to other materials and target properties can be found in reference [3].

### 7.2 Agent and Tools

The agent, powered by a large language model (LLM), is augmented with a set of tools allowing for a more chemistry-informed decision-making. These tools are as follows:

1. **Retrieval-Augmented Generation (RAG):** This tool allows the agent to obtain design guidelines on how to adapt the MOF structure from unstructured text. In specific, the agent has access to a fixed set of seven MOF research papers (see Refs. [4]-[10]) as PDFs. This tool is designed to extract the most relevant sentences from papers in response to a given query. It works by embedding both the paper and the query into numerical vectors, then identifying the top  $k$  passages within the document that either explicitly mention or implicitly suggest the adaptations to the band gap property for a MOF. The embedding model is OpenAI's text-ada-002 [11]. Inspired by our earlier work [12],  $k$  is set to 9 but is dynamically adjusted based on the relevant context's length to avoid OpenAI's token limitation error.
2. **Surrogate Band Gap Predictor:** The surrogate model used is a transformer (MOFormer [13]) that inputs the MOF as SMILES. This model is pre-trained using a self-supervised learning technique known as Barlow-Twin [14], where representation learning is done against structure-based embeddings from a crystal graph convolutional neural network (CGCNN) [15]. This was done against 16,000 BW20K entries [16]. The pre-trained weights are then transferred and fine-tuned to predict the band gap labels taken from 7450 entries from the QMOF database [17]. From a 5-fold training, an ensemble of five transformers are trained to return the mean band gap and the standard deviation, which is used to assess uncertainty for predictions. For comparison, our transformer's mean absolute error (MAE) is approximately 0.467, whereas MOFormer, which was pre-trained on 400,000 entries, achieves an MAE of approximately 0.387.
3. **Chemical Feasibility Evaluator:** This tool primarily uses RDKit [18] to convert a SMILES string into an RDKit Mol object, and performs several validation steps to ensure chemical feasibility. First, it parses the SMILES string to confirm correct syntax. Next, it validates the atoms and bonds, ensuring they are chemically valid and recognized. It then checks atomic valences to ensure each atom forms a reasonable number of bonds. For ring structures, RDKit verifies the correct ring closure notation. Additionally, it adds implicit hydrogens to satisfy valence requirements and detects aromatic systems,

marking relevant atoms and bonds as aromatic. These steps collectively ensure the molecule's basic chemical validity.

We use OpenAI's GPT-4 [19] with a temperature of 0.1 as our LLM and LangChain [20] for the application framework development (note the choice of LLM is only a hyperparameter and other LLMs can be also used with the agent).

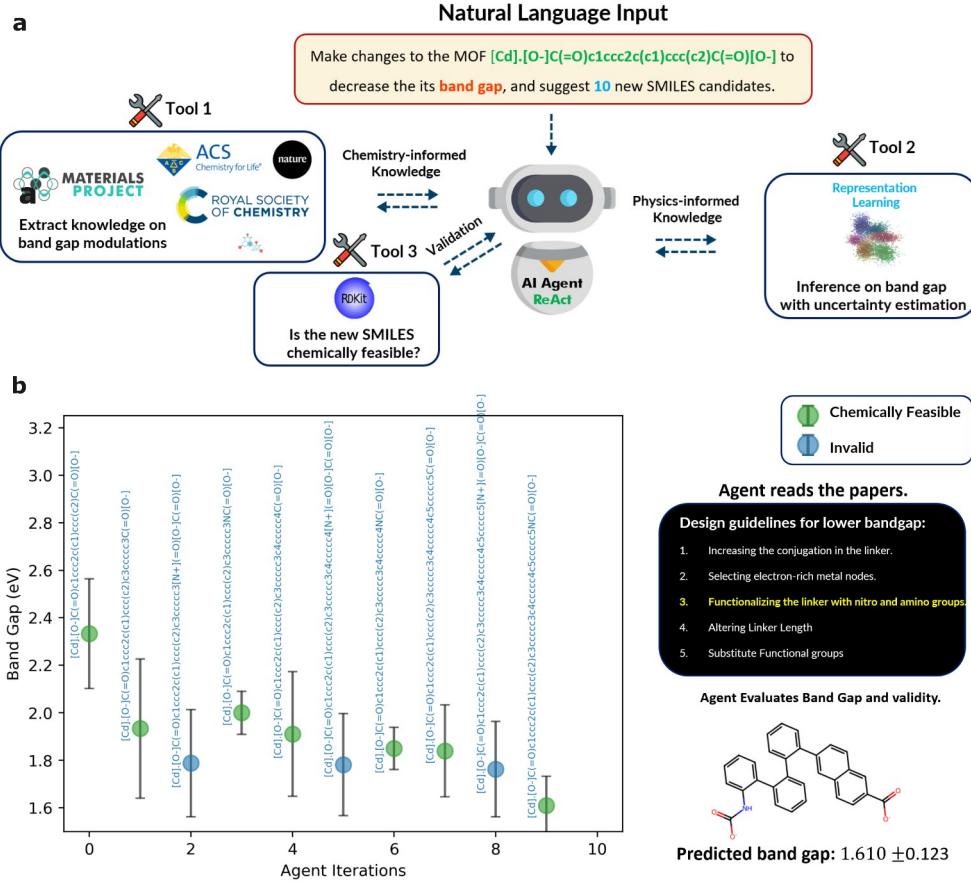


Figure 12: a) Workflow overview. The ReAct agent looks up guidelines for designing low band gap MOFs from research papers and suggests a new MOF (likely with lower band gap). It then checks validity of the new SMILES candidate and predicts band gap with uncertainty estimation using an ensemble of surrogate fine-tuned MOFormers. b) Band gap predictions for new MOF candidates as a function of agent iterations. Detailed analysis of this methodology applied to other materials and target properties can be found in reference [3].

The new MOF candidates and their corresponding inferred band gap are represented in Figure 1.b. The agent starts by retrieving the following design guidelines for low band gap MOFs from research papers: 1) Increasing the conjugation in the linker. 2) Selecting electron-rich metal nodes. 3) Functionalizing the linker with nitro and amino groups. 4) Altering linker length. 5) Substitute functional groups (i.e., substituting hydrogen with electron-donating groups on the organic linker). Note that the metal node adaptations were restrained by simply changing the system input prompt. The agent iteratively implements the above strategies and makes changes to the MOF. After each modification, the band gap of the new MOF is assessed using the fine-tuned surrogate MOFormers to ensure a lower band gap. Subsequently, the chemical feasibility is evaluated. If the new MOF candidate has an invalid SMILES string or a higher band gap, the agent reverts to the most recent valid MOF candidate with the lowest band gap.

### 7.3 Data and Code Availability

All code and data used to produce results in this study are publicly available in the following GitHub repository: <https://github.com/mehradans92/PoreVoyant>

## References

- [1] Lirong Li, Han Sol Jung, Jae Won Lee, and Yong Tae Kang. Review on applications of metal–organic frameworks for co<sub>2</sub> capture and the performance enhancement mechanisms. *Renewable and Sustainable Energy Reviews*, 162: 112441, 2022.
- [2] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. *arXiv preprint*, [arXiv:2210.03629](https://arxiv.org/abs/2210.03629), 2022.
- [3] Mehrad Ansari, Jeffrey Watchorn, Carla E. Brown, and Joseph S. Brown. dZiner: Rational Inverse Design of Materials with AI Agents. *arXiv*, 2410.03963, 2024. URL: <https://arxiv.org/abs/2410.03963>.
- [4] Muhammad Usman, Shruti Mendiratta, and Kuang-Lieh Lu. Semiconductor metal–organic frameworks: future low- g”bandgap materials. *Advanced Materials*, 29(6):1605071, 2017.
- [5] Espen Flage-Larsen, Arne Røyset, Jasmina Hafizovic Cavka, and Knut Thorshaug. Band gap modulations in uiO metal–organic frameworks. *The Journal of Physical Chemistry C*, 117(40):20610–20616, 2013.
- [6] Li-Ming Yang, Guo-Yong Fang, Jing Ma, Eric Ganz, and Sang Soo Han. Band gap engineering of paradigm mof-5. *Crystal growth & design*, 14(5):2532–2541, 2014.
- [7] Li-Ming Yang, Ponniah Vajeeston, Ponniah Ravindran, Helmer Fjellvag, and Mats Tilset. Theoretical investigations on the chemical bonding, electronic structure, and optical properties of the metal- organic framework mof-5. *Inorganic chemistry*, 49(22):10283–10290, 2010.
- [8] Maryum Ali, Erum Pervaiz, Tayyaba Noor, Osama Rabi, Rubab Zahra, and Minghui Yang. Recent advancements in mof-based catalysts for applications in electrochemical and photoelectrochemical water splitting: A review. *International Journal of Energy Research*, 45(2):1190–1226, 2021.
- [9] Yabin Yan, Chunyu Wang, Zhengqing Cai, Xiaoyuan Wang, and Fuzhen Xuan. Tuning electrical and mechanical properties of metal–organic frameworks by metal substitution. *ACS Applied Materials & Interfaces*, 15(36):42845–42853, 2023.
- [10] Chi-Kai Lin, Dan Zhao, Wen-Yang Gao, Zhenzhen Yang, Jingyun Ye, Tao Xu, Qingfeng Ge, Shengqian Ma, and Di-Jia Liu. Tunability of band gaps in metal–organic frameworks. *Inorganic chemistry*, 51(16):9039–9044, 2012.
- [11] Ryan Greene, Ted Sanders, Lilian Weng, and Arvind Neelakantan. New and improved embedding model, 2022.
- [12] Mehrad Ansari and Seyed Mohamad Moosavi. Agent-based learning of materials datasets from scientific literature. *arXiv preprint*, [arXiv:2312.11690](https://arxiv.org/abs/2312.11690), 2023.
- [13] Zhonglin Cao, Rishikesh Magar, Yuyang Wang, and Amir Barati Farimani. Moformer: self-supervised transformer model for metal–organic framework property prediction. *Journal of the American Chemical Society*, 145(5): 2958–2967, 2023.
- [14] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and St’ephane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.
- [15] Tian Xie and Jeffrey C. Grossman. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical Review Letters*, 120(14):145301, 2018.

- [16] Seyed Mohamad Moosavi, Aditya Nandy, Kevin Maik Jablonka, Daniele Ongari, Jon Paul Janet, Peter G Boyd, Yongjin Lee, Berend Smit, and Heather J Kulik. Understanding the diversity of the metal-organic framework ecosystem. *Nature communications*, 11(1):1–10, 2020.
- [17] Andrew S. Rosen, Shaelyn M. Iyer, Debmalya Ray, Zhenpeng Yao, Alan Aspuru-Guzik, Laura Gagliardi, Justin M. Notestein, and Randall Q. Snurr. Machine learning the quantum-chemical properties of metal-organic frameworks for accelerated materials discovery. *Matter*, 4(5):1578–1597, 2021.
- [18] Greg Landrum. RDKit documentation. Release, 1(1-79):4, 2013.
- [19] OpenAI. Gpt-4 technical report, 2023.
- [20] Harrison Chase. LangChain, 10 2022. URL <https://github.com/langchain-ai/langchain>.

## 8 How Low Can You Go? Leveraging SmallLLMs for Material Design

Authors: Alessandro Canalicchio, Alexander Moßhammer, Tehseen Rug, Christoph Völker

### 8.1 Motivation - Leveraging LLMs for Sustainable Construction Material Design

The construction industry's dependence on limited resources and the high emissions associated with traditional building materials such as Portland cement-based concrete necessitate a transition to more sustainable alternatives [1]. Geopolymers synthesized from industrial by-products such as fly ash and waste slag represent a promising solution [2-4]. However, scaling up these highly complex materials to meet market demand is a major challenge due to the smaller volumes available at diverse sources. Traditionally, it takes years of intensive scientific research to bring a new material to market. To meet the demand for sustainable alternative materials, new resource streams need to be developed much more frequently, making the traditional lengthy process of acquiring specialized material knowledge impractical.

This is where LLMs offer a solution. LLMs are trained on web-scale data and contain deep domain expertise that was previously hidden in scientific institutions but is now accessible to everyone. This knowledge is stored in the LLMs inner representations (hidden states) and is accessible through prompts making best practices from specialized domains easily available. In addition, LLMs can provide information in any format, making it easy to convert complex scientific concepts into workable formulations for direct application in the lab.

Previous systematic benchmarks have shown that LLMs can keep up with conventional approaches such as Bayesian Optimization (BO) and Sequential Learning (SL) when it comes to designing sustainable alternatives to concrete, so-called alkali-activated concrete [5]. What sets these models apart from classical design methods is their ability to produce zero-shot designs, meaning they can propose relatively well-functioning formulations without any initial training data. This capability suggests that LLMs can play a crucial role in the initial data collection phase. Essentially, it allows researchers entering a new field to begin new projects at an expert level from the outset, leading to much faster development of viable solutions.

### 8.2 Research Gap

The emergence of a new generation of high-performing, smaller-scale LLMs is expanding the boundaries of feasible application scenarios. Deploying these compact LLMs enables applications in sensitive areas like R&D. Full control over a local model enhances quality control and repeatability by allowing precise management of versions and system prompts. Additionally, they increase security and data privacy through in-house processing, which reduces the risk of confidential data leakage and other security threats. Moreover, small-scale LLMs improve cost efficiency and reliability by minimizing network dependency and operational costs. This development raises an intriguing research question:

Can current small LLMs retrieve and utilize valuable knowledge from their inner representations, making them effective tools for closed lab networks in material science?

Investigating this could significantly advance LLM adoption in material science R&D, especially in sensitive industry settings, potentially transforming experimental workflows and accelerating innovation.

### 8.3 Methodology

To investigate this question, we deployed two small-scale instruction LLMs on consumer hardware, specifically Llama 3 8B and Phi-3 3.8B, both quantized to four bits. Our goal was to evaluate their ability to design alkali-activated concrete with high compressive strength. The LLMs were tasked via system messages to determine four mixture parameters from a predetermined grid of 240 recipes: 1) the blending ratio of fly ash to ground granulated blast furnace slag, 2) the water-to-powder ratio, 3) the powder content, and 4) the curing method. We measured their performance by comparing the compressive strengths of the suggested formulations to previously published lab results [6].

The workflow and evaluation are shown in [Figure 13](#). Each design run comprised 10 consecutive development cycles, where the LLM suggested a formulation and the user provided feedback. This process was repeated five times for statistical analysis. Additionally, the prompt was rephrased synonymously three times to increase linguistic variability. Two types of contexts were provided: one with detailed design instructions, such as “reducing the water-to-powder ratio leads to higher strengths,” and one without additional instructions. In total, 15 design runs were conducted per model and per context. Finally, we assessed the achieved 10% lower-bound strength, defined as the strength achieved in 90% of the design runs, and compared this against a random draw as the statistical baseline and SL.

## 8.4 Results and Conclusion

The results, summarized in [Table 2](#), showed that all investigated models generated designs that outperformed the statistical baseline of 28 MPa in the first round and 52 MPa in the final round, demonstrating the surprising effectiveness of small-scale LLMs in materials design. An exception was noted with the Phi-3 model when provided with extensive design context: it produced the same output in each round and failed to produce viable solutions completely after the fifth development cycle, indicating its failure to understand the task. As expected, Llama 3 outperformed the smaller Phi-3 model. Specifically, Llama 3 benefited from additional design context, showing an improvement of more than 10 MPa in the initial rounds.

Table 2: Achieved compressive strength of designs suggested by LLMs. Statistical assessment in terms of 10% lower bound strength, i.e., the 10% worst cases.

Model	Development Cycle 1	Development cycle 10
Random Baseline	28 MPa	52 MPa
Phi 3 (No design rules in context)	50 MPa	58 MPa
Phi 3 (Design rules in context)	51 MPa	–
Llama 3 (No design rules in context)	48 MPa	60 MPa
Llama 3 (Design rules in context)	59 MPa	60 MPa

In conclusion, small-scale LLMs performed surprisingly well, with Phi-3 producing results significantly above a random guess, though it faced challenges with more complex prompts. The effectiveness of LLMs in solving design tasks depends on how well material concepts are represented in their hidden states and how effectively these can be retrieved via prompts, giving larger models an advantage. Despite their smaller parameter count and less training data, Phi-3 and Llama 3 demonstrated common sense for domain-specific design tasks, making local deployment a viable option. While 100% reliability in retrieving sensible information via LLMs is uncertain, small-scale LLMs can generate educated guesses that potentially accelerate the familiarization process with novel materials.

## 8.5 Code

The code used to conduct the experiments is open-source and available at <https://github.com/sandrocan/LLMs-for-design-of-alkali-activated-concrete-formulations>

## References

- [1] U. Environment, K. L. Scrivener, V. M. John and E. M. Gartner, "Eco-efficient cements: Potential economically viable solutions for a low-CO<sub>2</sub> cement-based materials industry," *Cement and Concrete Research*, vol. 114, pp. 2-26; DOI: <https://doi.org/10.1016/j.cemconres.2018.02.008>
- [2] J. L. Provis, A. Palomo and C. Shi, "Advances in understanding alkali-activated materials," *Cement and Concrete Research*, pp. 110-125, 2015.

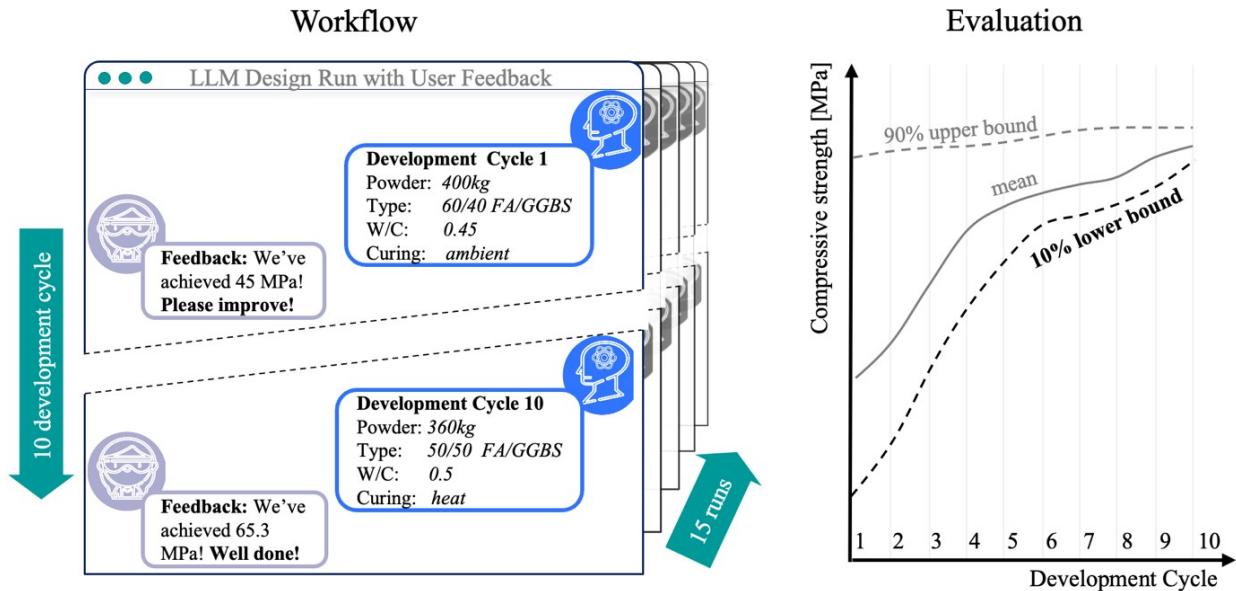


Figure 13: LLM-based material design workflow (left) and diagram showing evaluation metric (right).

- [3] H. S. Gölce, M. Tuyan, K. Ramyar and M. L. Nehdi, "Development of Eco-Efficient Fly Ash-Based Alkali-Activated and Geopolymer Composites with Reduced Alkaline Activator Dosage," *Journal of Materials in Civil Engineering*, vol. 32, no. 2, pp. 04019350; DOI: [10.1061/\(ASCE\)MT.1943-5533.0003017](https://doi.org/10.1061/(ASCE)MT.1943-5533.0003017) 2020.
- [4] J. He, Y. Jie, J. Zhang, Y. Yu and G. Zhang, "Synthesis and characterization of red mud and rice husk ash-based geopolymer composites," *Cement and Concrete Composites*, vol. 37, pp. 108-118; DOI: <http://dx.doi.org/10.1016/j.cemconcomp.2012.1201310>
- [5] C. Völker, T. Rug, K. M. Jablonbka and S. Kurschwitz, "LLMs can Design Sustainable Concrete – a Systematic Benchmark," (Preprint), pp. 1-12; DOI: [10.21203/rs.3.rs-3913272](https://doi.org/10.21203/rs.3.rs-3913272) 2023.
- [6] G. M. Rao and T. D. G. Rao, "A quantitative method of approach in designing the mix proportions of fly ash and GGBS-based geopolymer concrete," *Australian Journal of Civil Engineering*, vol. 16, no. 1, pp. 53-63; DOI: [10.1080/14488353.2018.1450208](https://doi.org/10.1080/14488353.2018.1450208) 2018.

## 9 LangSim

Authors: Yuan Chiang, Giuseppe Fisicaro, Greg Juhasz, Sarom Leang, Bernadette Mohr, Utkarsh Pratiush, Francesco Ricci, Leopold Talirz, Pablo A. Unzueta, Trung Vo, Gabriel Vogel, Sebastian Pagel, Jan Janssen

The complexity and non-intuitive user interface of scientific simulation software results in a high barrier for beginners and limits the usage to expert users. In the field of atomistic simulation, the simulation codes are developed by different communities (chemistry, physics, materials science) using different units, file names, and variable names. LangSim addresses this challenge, by providing a natural language interface for atomistic simulation in the field of computational materials science.

Since the introduction of ChatGPT, the application of large language models (LLM) in chemistry and materials science has transitioned from semantic literature search to research agents capable of autonomously executing selected steps of the research process. In particular, in research domains with a high level of automation, like chemical synthesis, the latest research agents already combine access to specialized databases, scientific software for analysis as well as to robots for executing the experimental measurements [1, 2]. These research agents divide the research question into a series of individual tasks, each addressing one task before combining them with one controlling agent. With this approach, the flexibility of the LLM is reduced, which consequently reduces the risk of hallucinations [8].

In analogy, LangSim (Language + Simulation) is a research agent focused on simulation in the field of computational materials science. LangSim can calculate a series of bulk properties for elemental crystals, like the equilibrium volume and equilibrium bulk modulus. Internally, this is achieved by constructing simulation protocols consisting of multiple simulation steps to calculate one material property. For example, the bulk modulus is calculated by querying the atomistic simulation environment (ASE) [3] for the equilibrium crystal structure, optimizing the crystal structure in dependence on the choice of simulation model, and finally evaluating the change of energy over volume change around the equilibrium to calculate the bulk modulus as the second derivative of the change in energy over volume change. The simulation protocols in LangSim are independent of the selected level of theory and can be evaluated with either the effective medium theory model [4] or the foundation machine-learned interatomic potential MACE [5]. Furthermore, to quantify the uncertainty of these simulation results, LangSim also has access to databases with experimental references for these bulk properties.

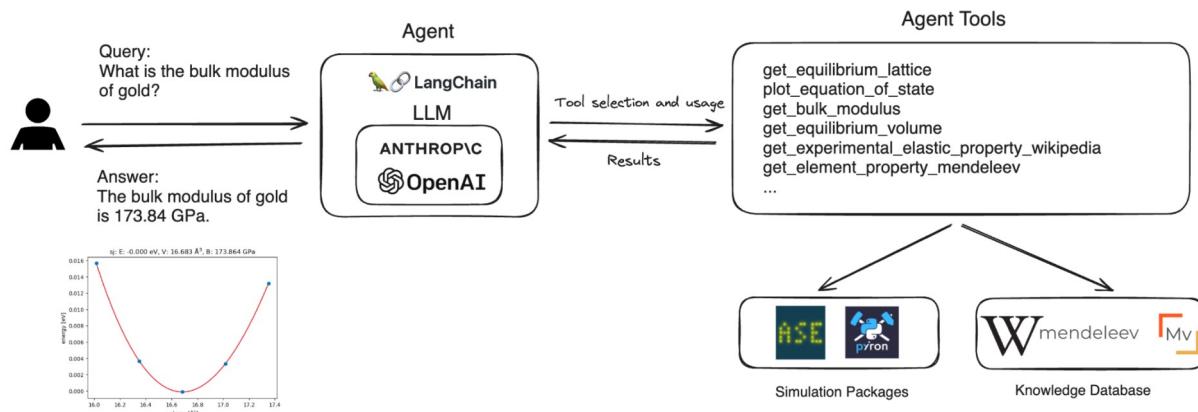


Figure 14

The LangSim research agent is based on the LangChain package. This has two advantages: On the one hand, the LangChain package [6] simplifies the addition of new simulation agents and simulation workflows for LangSim. On the other hand, LangChain is compatible with a wide range of different LLM providers to prevent vendor lock-in. LangSim extends the LangChain framework by providing data types for coupling the simulation codes with the LLM, like a pydantic dataclass [7] representation of the ASE atoms class and a series of pre-defined simulation workflows to highlight how existing simulation workflows can be implemented.

as LangChain agents. Once a simulation workflow is represented as a Python function compatible with a simulation framework like ASE, the interfacing with LangSim is as simple as changing the input arguments to LLM-compatible data types indicated by type hints and adding a Docstring as context for the LLM.

Abstractly, these LangChain agents can be understood in analogy to the header files in C programming, which define the interfaces for public functions. An example LangSim agent to calculate the bulk modulus is provided below:

```
from ase.atoms import Atoms
from ase.calculators import Calculator
from ase.eos import calculateeos
from ase.units import kJ
from langsim import (
    AtomsDataClass,
    getasecalculatorfromstr,
)
from langchain.agents import tool

def getbulkmodulusfunction(
    atoms: Atoms, calculator: Calculator
) -> float:
    atoms.calc = calculator
    eos = calculateeos(atoms)
    v, e, B = eos.fit()
    return B / kJ * 1.0e24

@tool
def getbulkmodulusagent(
    atomdict: AtomsDataClass, calculatorstr: str
) -> float:
    """
    Returns the bulk modulus of chemical symbol
    for a given atoms dictionary and a selected
    model specified by the calculator string in GPa.
    """
    return getbulkmodulusfunction(
        atoms=Atoms(**atomdict.dict()),
        calculator=getasecalculatorfromstr(
            calculatorstr=calculatorstr
        ),
    )
```

The example workflow for calculating the bulk modulus highlights how existing simulation frameworks like ASE can be leveraged to provide the LLM with the ability to construct and execute simulation workflows. While for this example, in particular for the case of elemental crystals, it would be possible to pre-compute all combinations and restrict the large language model to a database of pre-computed results, this would become prohibitive for multi-component alloys and an increasing number of simulation models and workflows. At this stage the semantic capabilities of the LLM provide the capability to handle all possible combinations in a systematic way by allowing the LLM to construct and execute the specific simulation workflow when it is requested from the user.

In summary, the LangSim package provides data classes and utility functions to interface LLMs with atomistic simulation workflows. This provides the LLM with the ability to execute simulations to answer scientific questions like a computational material scientist. The functionality is demonstrated for the calculation of the bulk modulus for elemental crystals.

## 9.1 One sentence summaries

1. Problem/Task: Develop a natural language interface for simulation codes in the field of computational chemistry and materials science. Current LLMs, including ChatGPT 4, suffer from hallucination, resulting in simulation protocols that can be executed but fail to calculate the correct physical property with the specified unit.
2. Approach: Develop a suite of LangChain agents to interface with the atomic simulation environment (ASE) and corresponding data classes to represent objects used in atomistic simulation in the context of large language models.
3. Results and Impact: Developed the LangSim package as a prototype for handling the calculation of multiple material properties using predefined simulation workflows, independent of the theoretical model, based on the ASE framework.
4. Challenges and Future Work: The current prototype enables the calculation of bulk properties for unaries, the next step is to extend this functionality to multi-component alloys and more material properties.

## References

- [1] Boiko, D.A., MacKnight, R., Kline, B. et al. Autonomous chemical research with large language models. *Nature* 624, 570–578 (2023). <https://doi.org/10.1038/s41586-023-06792-0>
- [2] M. Bran, A., Cox, S., Schilter, O. et al. Augmenting large language models with chemistry tools. *Nat Mach Intell* 6, 525–535 (2024). <https://doi.org/10.1038/s42256-024-00832-8>
- [3] Hjorth Larsen, A., Jørgen Mortensen, J., Blomqvist, J., Castelli, I. E., Christensen, R., Dulak, M., Friis, J., Groves, M. N., Hammer, B., Hargus, C., Hermes, E. D., Jennings, P. C., Bjerre Jensen, P., Kermode, J., Kitchin, J. R., Leonhard Kolsbjerg, E., Kubal, J., Kaasbjerg, K., Lysgaard, S., . . . Jacobsen, K. W. (2017). The atomic simulation environment—a python library for working with atoms. *Journal of Physics: Condensed Matter*, 29(27), 273002. <https://doi.org/10.1088/1361-648x/aa680e>
- [4] Jacobsen, K. W., Stoltze, P., Nørskov, J. K. (1996). A semi-empirical effective medium theory for metals and alloys. *Surface Science*, 366(2), 394–402. [https://doi.org/10.1016/0039-6028\(96\)00816-3](https://doi.org/10.1016/0039-6028(96)00816-3)
- [5] Batatia, I., Benner, P., Chiang, Y., Elena, A. M., Kováč, D. P., Riebesell, J., Advincula, X. R., Asta, M., Avaylon, M., Baldwin, W. J., Berger, F., Bernstein, N., Bhowmik, A., Blau, S. M., Cárare, V., Darby, J. P., De, S., Della Pia, F., Deringer, V. L. et al. (2024). A foundation model for atomistic materials chemistry. *arXiv*. <https://arxiv.org/abs/2401.00096>
- [6] <https://github.com/langchain-ai/langchain>
- [7] <https://pydantic.dev/>
- [8] Ye, H., Liu, T., Zhang, A., Hua, W., Jia, W. (2023). Cognitive mirage: A review of hallucinations in large language models. *arXiv*. <https://arxiv.org/abs/2309.06794>

## 10 LLMicroscopilotassisting microscope operations through LLMs

*Authors: Marcel Schloz, Jose C. Gonzalez*

The operation of state-of-the-art microscopes in materials science research is often limited to a selected group of operators due to their high complexity and significant cost of ownership. This exclusivity creates a barrier to broadening scientific progress and democratizing access to these powerful instruments. Presently, operating these microscopes involves time-consuming tasks that demand substantial human expertise, such as aligning the instrument for optimal performance and transitioning between different operational modes to address different research questions. These challenges highlight the need for improved user interfaces that simplify operation and increase the accessibility of microscopes in materials science.

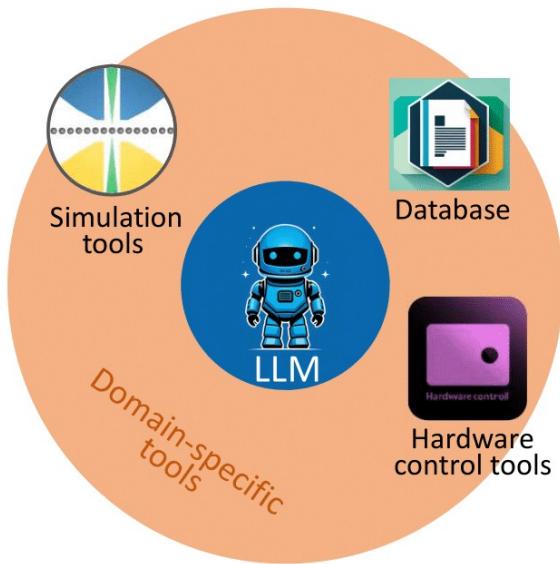
Recent advancements in natural language processing software suggest that integrating large language models (LLMs) into the user experience of modern microscopes could significantly enhance their usability. Just as modern chatbots have enabled users without much programming background to create complex computer programs, LLMs have the potential to simplify the operation of microscopes, thereby making them more accessible to non-expert users [1]. Early studies have demonstrated the potential of LLMs in scanning probe microscopy, using microscope-specific external tools for remote access [2] and control [3]. Particularly promising is the application of LLMs as agents with access to specific external tools, providing operators with a powerful assistant capable of reasoning based on observations and reducing the extensive hallucinations common in LLM agents. This approach also enhances the accessibility of external tools, eliminating the need for users to learn tool-specific APIs.

The LLMicroscopilot-team (Jose D. Cojal Gonzalez and Marcel Schloz) has shown that the operation of a scanning transmission electron microscope can be partially performed by the LLM-powered agent "LLMicroscopilot" through access to microscope-specific control tools. Figure 15 illustrates the interaction process between the operator and the LLMicroscopilot. The LLMicroscopilot is built on a generally trained foundation model that gains domain-specific knowledge and performance through the provided tools. The initial prototype uses the API of a microscope experiment simulation tool [4] to perform tasks such as experimental parameter estimation and experiment execution. This approach reduces the reliance on highly trained human operators, fostering broader participation in materials science research. Future developments of LLMicroscopilot will integrate open-source microscope hardware control tools [5] and database-access tools, allowing for Retrieval-Augmented Generation possibilities to improve parameter estimation and data analysis.

## References

- [1] Stefan Bauer et al, Roadmap on data-centric materials science, *Modelling Simul. Mater. Sci. Eng.*, 2024, 32, 063301.
- [2] Diao, Zhuo, Hayato Yamashita, and Masayuki Abe. "Leveraging Large Language Models and Social Media for Automation in Scanning Probe Microscopy." *arXiv preprint arXiv:2405.15490* (2024).
- [3] Liu, Yongtao, Marti Checa, and Rama K. Vasudevan. "Synergizing Human Expertise and AI Efficiency with Language Model for Microscopy Operation and Automated Experiment Design." *Machine Learning: Science and Technology* (2024).
- [4] Madsen, Jacob, and Toma Susi. "The abTEM code: transmission electron microscopy from first principles." *Open Research Europe* 1 (2021).
- [5] Meyer, Chris, et al. "Nion Swift: Open Source Image Processing Software for Instrument Control, Data Acquisition, Organization, Visualization, and Analysis Using Python." *Microscopy and Microanalysis* 25.S2 (2019): 122-123.

## LLMicroscopilot



## Microscope User Interface

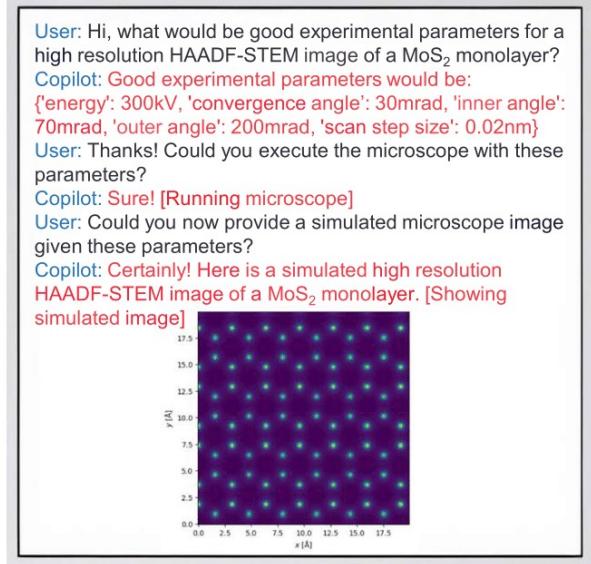


Figure 15: Schematic overview of the LLMicroscopilot assistant. The microscope user interface allows the user to input queries, which are then processed by the LLM. The LLM executes appropriate tools to provide domain-specific knowledge, support data analysis, or operate the microscope.

# 11 T2Dllama: Harnessing Language Model for Density Functional Theory (DFT) Parameter Suggestion

Authors: Chiku Parida, Martin H. Petersen

## 11.1 Introduction

Large language models are now gaining the attention of many researchers due to their capabilities to process human language and perform tasks on which they have not been explicitly trained, making them an invaluable tool for researchers in various fields where the information is in the form of text, like scientific journals, blogs, news articles, and social media posts, etc. This is particularly applicable to the field of chemical sciences, which encounters the challenge of dealing with limited and diverse datasets that are often presented in text format. LLMs have proven their potential in handling these challenges and are progressively being utilised to predict chemical characteristics, optimise reactions, and even independently design and execute experiments [1].

Here, we used LLM to process published scientific articles to extract simulation parameters and other relevant information about different materials. This will help experimentalists get an idea of optimised parameters for Density Functional Theory (DFT) calculations for the newly discovered material family.

Nowadays, DFT is the most valuable tool to model atomistic materials. The idea behind DFT is to use Kohn-Sham's equations to approximately solve Schrödinger's equations for the atomic material at hand. The approximation is done by configuring the electron density for the material at each ionic step, where the ions move in position based on their energy and forces determined by the configured electron density. The biggest part of the approximation is the exchange functional, and depending on the complexity of the exchange functional, the approximation becomes more or less comparable with experimental results [6]. When performing a DFT calculation, the question is always what exchange functional and parameters to use, as well as what k-space grid to use. This is material-dependent and will change for different materials. The unoptimized parameters can lead to inaccurate results, resulting in a DFT calculation that fails to describe the relative values and is therefore not comparable to the experimental results [2]. For that reason, experimentalists normally collaborate with computational chemists because of their expertise in computational modeling or make due without the atomistic model. Instead, our T2Dllama [talk-to-documents using Llama] framework can be an acceptable solution to give the necessary DFT parameters, and using additional tools on the top of the LLM interface, it can create inputs for atomistic simulations [3, 4] from the provided structure file by the user.

## 11.2 Retrieval Augmented Generation

The retrieval augmented generation (RAG) technique [5] is a popular and efficient technique for generating relevant contextual responses using pre-trained models. It is also less complex as compared to fine-tuning and training LLM from scratch. Figure 16 explains our RAG workflow. The RAG workflow has the following blocks:

*Data Preparation:* We collect open-access scientific journals using Arxiv-api and the necessary filters. Pre-existing licensed scientific journals in the local database can be used for confidential purposes. Then the text documents are processed to create chunks of text.

*Indexing and Embedding:* In this crucial step, we use llama indexing [7] to store the tokenized documents with vector indexes and embeddings, transforming them into knowledge vector databases.

*Information Retrieval:* When the LLM interface receives a human prompt, the model searches the knowledge vector database and retrieves the relevant chunks of stored data.

*LLM Interface:* This is the last step where we communicate with the user. The LLM interface received the prompt from the user and got the information, as explained previously. The retrieved chunks during information retrieval are processed by the pretrained model [Mistral 7B] [8], and the created context is delivered to the user as a response.

## Retrieval Augmented Generation [RAG]

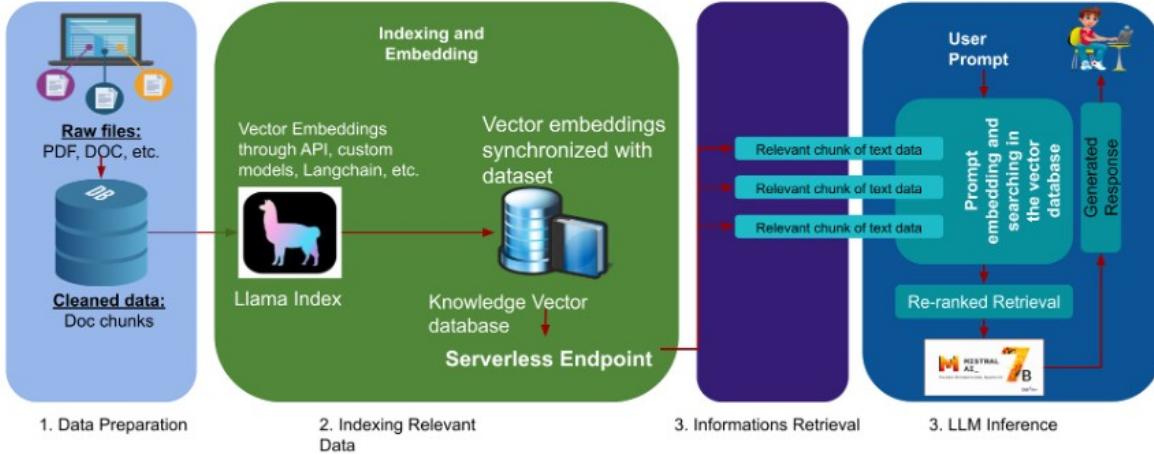


Figure 16: Retrieval Augmented Generation [RAG] architecture with LLM interface

### 11.3 Summary and outlooks

The biggest issue here is training a GPT model to predict DFT exchange functionals and parameters is filtering the data. There is no consensus about using a specific exchange functionals and parameters for a material, which makes the model confused. A way to avoid this is to only use papers, where DFT calculations are directly comparable with experimental results. This will limit the variability in the data and ensure that the model is trained on reliable information. To conclude through a pre-trained GPT model, we are able to predict DFT exchange functionals and Hubbard-U parameters as well as k-point grid by using RAG technique. Further, as we are retrieving information about simulation parameters and exchange-functionals for materials consistent with respect to experiments from the documents, it is important to prioritise relevant and reliable sources and special techniques to ensure the accuracy of the data being extracted. One approach that can be taken involves employing advanced tokenizer techniques specifically designed for scientific notations and terms, so that we can get reliable embedding and vector indexing. This will help to improve the quality of the responses. We still need to do more thorough training as well as a GUI application, but as an initial try in this LLM hackathon we showed that the idea can be realised.

### 11.4 Data and Code Availability

All code and data used in this study are publicly available in the following GitHub repository: <https://github.com/chiku-parida/T2DlLama>

## References

- [1] Adrian Mirza et al., "Are large language models superhuman chemists?" <https://doi.org/10.48550/arXiv.2404.01475>
- [2] Hafner, Jürgen, "Ab-initio simulations of materials using VASP: Density-functional theory and beyond." *Journal of computational chemistry* 29.13 (2008): 2044-2078.
- [3] Kresse, Georg, and Jürgen Hafner, "Ab initio molecular dynamics for liquid metals." *Physical review B* 47.1 (1993): 558.

- [4] Mortensen, Jens Jørgen, Lars Bruno Hansen, and Karsten Wedel Jacobsen, "Real-space grid implementation of the projector augmented wave method." *Physical Review B—Condensed Matter and Materials Physics* 71.3 (2005):035109.
- [5] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bic, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang, "Retrieval-Augmented Generation for Large Language Models: A Survey", <https://doi.org/10.48550/arXiv.2312.10997>
- [6] Giustino, Feliciano. *Materials modelling using density functional theory: properties and predictions*. Oxford University Press, 2014.
- [7] Llamaindex, [https://docs.llamaindex.ai/en/stable/examples/llm/llama\\_2\\_llama\\_cpp/](https://docs.llamaindex.ai/en/stable/examples/llm/llama_2_llama_cpp/)
- [8] Mistral 7B, the model used, <https://huggingface.co/TheBloke/Mistral-7B-Instruct-v0.1-GGUF>

## 12 Materials AgentAn LLM-Based Agent with Tool-Calling Capabilities for Cheminformatics

Authors: Archit Datar, Kedar Dabhadkar

### 12.1 Introduction

Out-of-the box large language model (LLM) implementations such as ChatGPT, while offering interesting responses, generally provide little to no control over the workflow of the LLM by which the response is generated. In other words, it is easy to get LLMs to say something in response to a prompt, but difficult to get them to do something via an expected workflow. A solution to do this problem is to equip LLMs with tool-calling capabilities; i.e., allow the LLM to generate the response via an available function(s) which is (are) appropriate to answer the prompt. An LLM with tool-calling capabilities, when prompted, typically decides which tool(s) to call (execute) and the order in which to execute them, along with the arguments to pass to them. It then executes these and returns the response. Such a system offers several powerful capabilities such as the ability to query databases to return the latest information and reliably perform mathematical calculations. Such capabilities have been incorporated into ChatGPT via plugins such as Expedia and Wolfram, among others [1]. In the chemistry literature, recent attempts have been made by researchers such as Smit and coworkers, Schwaller and coworkers, among others [2, 3].

Through Materials Agent, which is an LLM-based agent with tool-calling capabilities for cheminformatics, we seek to build on these attempts to provide a variety of important tool-calling capabilities and build a framework to expand on these. We hope that this can serve to increase LLM adoption in the community and lower the barrier to entry for cheminformatics. Materials Agent is built using the LangChain library [4], GPT 3.5-turbo [5] as the underlying LLM, and the user interface is based on the FastDash project [6]. In this demonstration, we have provided tools based on RDKit [7]—a popular cheminformatics library, some custom tools, as well as a Retrieval Augmented Generation (RAG) to allow LLM to interact with documents. The full code is available at <https://github.com/dkedar7/materials-agent> and the working application, hosted on Google Cloud Platform (GCP), is available at <https://materials-agent-hpn4y2dvda-ue.a.run.app/>. The demonstration video is uploaded on YouTube at [https://www.youtube.com/watch?v=\\_5yC0g5Bi\\_Q&ab\\_channel=ArchitDatar](https://www.youtube.com/watch?v=_5yC0g5Bi_Q&ab_channel=ArchitDatar). In the following section, we describe some key use cases.

### 12.2 Equipping LLMs with tools based on standard cheminformatics packages (RDKit)

Common cheminformatics workflows involve obtaining SMILES strings and key properties of molecules. Searching for these data individually can be time consuming. The RDKit CalcMolDescriptors module comes pre-loaded with these data based on SMILES strings. We created a function to select 19 of the most common properties (for ease of visualization) and added it as a tool to the LLM. Not only can the LLM perform the simple task of providing these descriptions when a SMILES string is input, but it can also perform more complicated tasks involving this functionality. As shown in Figure 17 below, given a complex query, the LLM breaks it down into individual tasks, utilizes these tools in the correct sequence, and renders a response.

### 12.3 Tool-calling for custom tools: Radial Distribution Function calculation

Combining custom tools with LLMs lead to some interesting advantages. For one, they offer the makers of these custom tools an ability to provide their users with a more intuitive natural language-based user interface which can lower the barrier to entry and increase adoption. Other benefits can be that they can be integrated into other workflows and automate more work. To demonstrate this, we have built a tool which computes and plots the radial distribution function (RDF) and integrated it with an LLM. The RDF is an important function commonly used in molecular dynamics and Monte Carlo simulations to understand the statistics of distances between two particles averaged over the simulation. By studying these, one can understand the nature of interactions between these particles. Here, we constructed a custom tool to compute RDF for the distance between a water molecule and the framework atoms in a metal-organic framework (MOF) in a single molecule canonical ensemble Monte Carlo simulation. The inputs are a PDB file of the MOF structure and

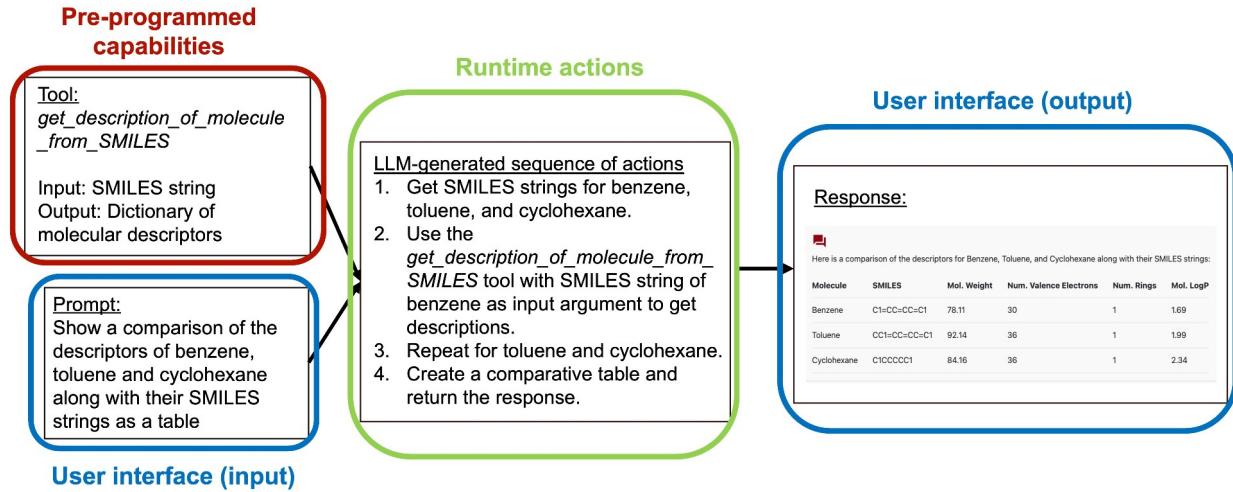


Figure 17: Workflow of a response generated to a user prompt by Materials Agent using a tool based on RDKit.

a TXT file containing snapshots of the location of the water molecule during the simulation. The distance computation also accounts for triclinic periodic boundary conditions which is the accurate way to quantify distances for crystalline systems such as this one. The inputs and outputs for this tool are shown below in Figure 18(a).

Furthermore, we also stress that this approach is easily scalable and transferable, and adding new tools is exceedingly easy. The reader is encouraged to clone our GitHub repository and experiment with adding new tools to this software package. New tools can be easily added to the `src/tools.py` in the repository via the format shown in the code snippet in Figure 18(b).

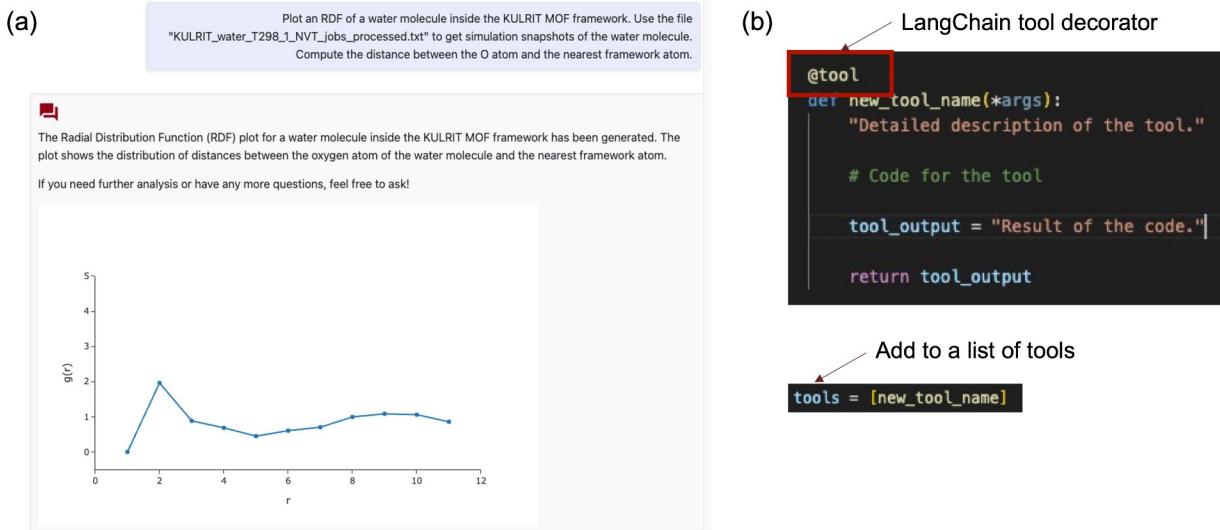


Figure 18: Niche purpose tools with LLM. (a) Illustration of the RDF computing tool. (b) Code snippet to highlight the ease of transferability of LLMs with tool-calling capabilities.

## 12.4 RAG capabilities

Summarizing and asking questions of documents is another common LLM use case. This capability is provided out-of-the box through the EmbedChain library [8]. and we have integrated that into Materials Agent for convenience. We demonstrate the utility of this by supplying the LLM with a URL to a materials and safety datasheet (MSDS) and asking questions of it (see Figure 19).

**Materials Agent**



AGENT

Can you please read this MSDS from this URL? <https://www.airgas.com/msds/001029.pdf>. What are the main hazards associated with this material?

The main hazards associated with the material specified in the MSDS from the provided URL include:

- Extremely flammable gas
- Contains gas under pressure that may explode if heated
- Fatal if inhaled
- Causes respiratory irritation
- Very toxic to aquatic life

If you have more questions or need further information, feel free to ask!

What are the synonyms of this material?

The alternate names or synonyms for the material specified in the MSDS from the provided URL are:

- Hydrogen sulfide
- Sulfuretted hydrogen
- Sewer gas
- Hydrosulfuric acid
- Dihydrogen sulfide
- Hydrosulfuric acid
- Sulphuretted hydrogen
- Sulfur hydride
- Hydrogen sulfide (H2S)
- Sulfuric acid Gas

If you have any more questions or need further assistance, feel free to ask!

**SAFETY DATA SHEET**

**Airgas.**  
an Air Liquide company

**Section 1. Identification**

**GHS product identifier**: Hydrogen Sulfide

**Chemical name**: hydrogen sulfide

**Other means of identification**: Hydrogen sulfide; Sulfuretted hydrogen; Sewer gas; Hydrosulfuric acid; dihydrogen sulfide; hydrosulfuric acid; sulphuretted hydrogen; HYDROGEN SULFIDE H2S; Sulfur hydride; Hydrogen sulfide (H2S); Sulfuric acid

**Product type**: Gas

**Product use**: Synthetic/Analytical chemistry

**Synonym**: Hydrogen sulfide; Sulfuretted hydrogen; Sewer gas; Hydrosulfuric acid; dihydrogen sulfide; hydrosulfuric acid; sulphuretted hydrogen; HYDROGEN SULFIDE H2S; Sulfur hydride; Hydrogen sulfide (H2S); Sulfuric acid

**SDS #**: 001029

**Supplier's details**: Airgas USA, LLC and its affiliates  
2900 Radnor Radnor-Chester Road  
Suite 100  
Radnor, PA 19087-5283  
1-810-687-5253

**24-hour telephone**: 1-866-734-3438

**Section 2. Hazards identification**

**OSHA/HCS status**: This material is considered hazardous by the OSHA Hazard Communication Standard (29 CFR 1910.1200).

**Classification of the substance or mixture**: FLAMMABLE GASES - Category 1  
GASES UNDER PRESSURE - Liquefied gas  
ACUTE TOXICITY (inhalation) - Category 2  
SPECIFIC TARGET ORGAN TOXICITY (SINGLE EXPOSURE) (Respiratory tract irritation) - Category 3  
AQUATIC HAZARD (ACUTE) - Category 1

**GHS label elements**

**Hazard pictograms**:

**Signal word**: Danger

**Hazard statements**: Extremely flammable gas. Contains gas under pressure; may explode if heated. Fatal if inhaled. May cause respiratory irritation. Very toxic to aquatic life. Extended exposure to gas reduces the ability to smell sulfides. May form explosive mixtures with air.

**Precautionary statements**

**General**: Read and follow all Safety Data Sheets (SDS's) before use. Read label before use. Keep out of reach of children. If medical advice is needed, have product container or label at hand. Close valve after each use and when empty. Use equipment rated for cylinder pressure. Do not open valve until connected to equipment prepared for use. Use a back flow preventative device in the piping. Use only equipment of compatible materials of construction. Always keep container in upright position. Do not depend on odor to detect presence of gas. Approach suspected leak area with caution.

Date of issue/Date of revision : 2/11/2021 Date of previous issue : 1/13/2017 Version : 1.01 1/12

Figure 19: Illustration of RAG for interacting with a MSDS.

In future, we aim to expand the toolkit that the LLM is equipped with. For instance, we can add functions built on the publicly available PubChem database [9]. as well as some functions built off of it [10]. We also aim to train it on user manuals of commonly used molecular simulations software such as GROMACS [11], RASPA [12], and QuantumEspresso [13] to assist with setting up molecular simulations.

Through the experience of building Materials Agent, we realized that, while convenient, such agents cannot replace the need for human vigilance. At the same time, having the development of such an agent will make cheminformatics utilities easier to access for a broader range of users, lower the barrier to entry, and ultimately, accelerate the pace of materials development.

## References

- [1] OpenAI plugins, <https://openai.com/index/chatgpt-plugins/>
- [2] Jablonka, K.M., Schwaller, P., Ortega-Guerrero, A. et al. Leveraging large language models for predictive chemistry. *Nat Mach Intell* 6, 161–169 (2024). <https://doi.org/10.1038/s42256-023-00788-1>
- [3] M. Bran, A., Cox, S., Schilter, O. et al. Augmenting large language models with chemistry tools. *Nat Mach Intell* 6, 525–535 (2024). <https://doi.org/10.1038/s42256-024-00832-8>
- [4] LangChain, <https://www.langchain.com/>

- [5] OpenAI models, <https://platform.openai.com/docs/models>
- [6] Fast Dash, <https://docs.fastdash.app/>
- [7] RDKit: Open-source cheminformatics;<http://www.rdkit.org>
- [8] Singh, Taranjeet. Embedchain, <https://github.com/embedchain/embedchain>
- [9] PubChem programmatic access,<https://pubchem.ncbi.nlm.nih.gov/docs/programmatic-access>
- [10] PubChemPy, <https://pubchempy.readthedocs.io/en/latest/guide/introduction.html>
- [11] Abraham, M., Alekseenko, A., Basov, V., Bergh, C., Briand, E., Brown, A., Doijade, M., Fiorin, G., Fleischmann, S., Gorelov, S., Gouaillardet, G., Grey, A., Irrgang, M. E., Jalalypour, F., Jordan, J., Kutzner, C., Lemkul, J. A., Lundborg, M., Merz, P., . . . Lindahl, E. (2024). GROMACS 2024.2 Manual (2024.2). Zenodo. <https://doi.org/10.5281/zenodo.11148638>
- [12] Dubbeldam, D., Calero, S., Ellis, D. E., & Snurr, R. Q. (2015). RASPA: molecular simulation software for adsorption and diffusion in flexible nanoporous materials. *Molecular Simulation*, 42(2), 81–101. <https://doi.org/10.1080/08927022.2015.1010082>
- [13] Giannozzi, P., Baroni, S., Bonini, N., Calandra, M., Car, R., Cavazzoni, C., Ceresoli, D., Chiarotti, G. L., Cococcioni, M., Dabo, I., Dal Corso, A., de Gironcoli, S., Fabris, S., Fratesi, G., Gebauer, R., Gerstmann, U., Gougaussis, C., Kokalj, A., Lazzeri, M., . . . Wentzcovitch, R. M. (2009). Quantum Espresso: A modular and open-source software project for quantum simulations of materials. *Journal of Physics: Condensed Matter*, 21(39), 395502. <https://doi.org/10.1088/0953-8984/21/39/395502>

## 13 LLM with Molecular Augmented Token

Authors: Luis Pinto, Xuan Vu Nguyen, Tirtha Vinchurkar, Pradip Si, Suneel

Kuman

### 13.1 Objective

Our primary objective is to explore how chemical encoders such as molecular fingerprints or embeddings from 2D/3D deep learning models (e.g., ChemBERTa [1], UniMol [2]) can enhance large language models (LLMs) for zero-shot tasks such as property prediction, molecule editing, and generation. We aim to benchmark our approach against state-of-the-art models like LlaSmol [3] and ChatDrug [4], demonstrating the transformative potential of LLMs in the field of chemistry.

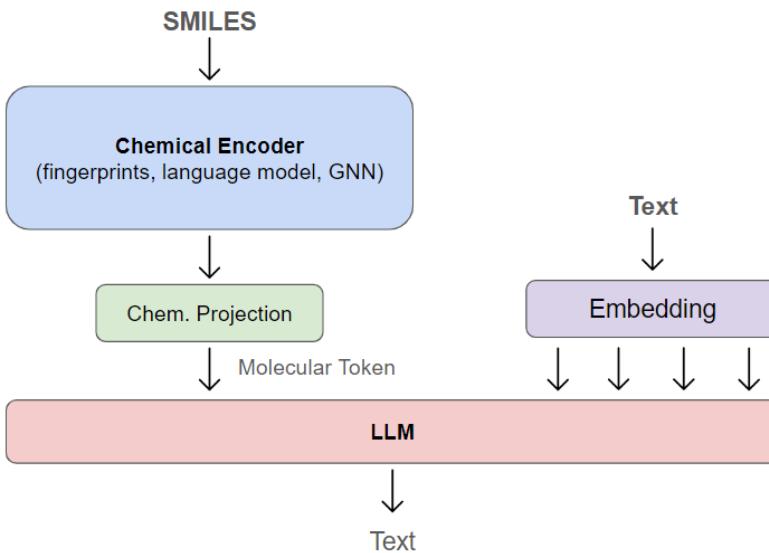


Figure 20: Workflow for integrating chemical encoders with large language models. Molecular data from SMILES is transformed into molecular tokens and combined with text embeddings for tasks such as property prediction, molecule editing, and generation.

### 13.2 Methodology

We identified two key benchmarks to evaluate our approach:

- **LlaSmol:** This benchmark involves fine-tuning a Mistral 7B model [5] on 14 different chemistry tasks, including 6 property prediction tasks. The LlaSmol project demonstrated significant performance improvements over baseline models, both open-source and proprietary, by using the SMolInstruct dataset, which contains over three million samples.
- **ChatDrug:** This framework leverages ChatGPT for molecule generation and editing tasks. It includes a prompt module, a retrieval and domain feedback module, and a conversation module to facilitate effective drug editing. ChatDrug showed superior performance across 39 drug editing tasks, encompassing small molecules, peptides, and proteins, and provided insightful explanations to enhance interpretability.

### 13.3 Our approach

We propose to fine-tune a Mistral 7B instruct model to compete against these benchmarks. Due to compute constraints, we were unable to complete the training. Training the model using QLoRA 4-bit [6] on 50k property prediction samples requires approximately 5 hours per epoch on a 24GB GPU.

Steps Taken:

- **Data Preparation:** We utilized chemical encoders to transform molecular structures into suitable embeddings for the LLM.
- **Model Modification:** We integrated the embeddings into the LLM's forward function to enrich its input.
- **Fine-Tuning:** We applied QLoRA for efficient training on limited computational resources.
- **Preliminary Results and Ongoing Work:** Although we are still fine-tuning the LLMs, initial results are promising.
- **Code Snippets:** Screenshots in Figures 21 and 22 demonstrate the modifications made to the model code to extract embeddings and implement the forward function of the LLM.

### 13.4 Conclusion

Our project underscores the potential of using LLMs enhanced with chemical encoders in materials science and chemistry. By fine-tuning these models, we aim to improve property prediction and facilitate molecule editing and generation, paving the way for future research and applications in this space. The code is available at <https://github.com/luispintoc/LLM-mol-encoder>

```
def get_embeddings(self, smiles_tokens, text_tokens):  
    # Encoder forward pass / Get SMILES embeddings  
    smiles_tokens = {k: v.to('cuda') for k, v in smiles_tokens.items()}  
    text_tokens = {k: v.to('cuda') for k, v in text_tokens.items()}  
  
    mol_encoder_output = self.mol_encoder(**smiles_tokens)  
    smiles_embedding = mol_encoder_output['last_hidden_state'][:, 0, :]  
    smiles_projection = self.linear_project(smiles_embedding).unsqueeze(1)  
  
    # Get embeddings from LLM for the question  
    embedding_layer = self.llm_model.model.model.embed_tokens  
    llm_embeddings = embedding_layer(text_tokens['input_ids']).squeeze(1)  
  
    # Concatenate encoder and LLM embeddings  
    combined_embeddings = torch.cat((smiles_projection, llm_embeddings), dim=1)  
  
    return combined_embeddings
```

Figure 21: Modified code of the function get embeddings.

## References

- [1] S. Chithrananda, G. Grand, and B. Ramsundar, "ChemBERTa: Large-Scale Self-Supervised Pretraining for Molecular Property Prediction," *arXiv preprint arXiv:2010.09885*, 2020. Available: <https://arxiv.org/abs/2010.09885>
- [2] G. Zhou, Z. Gao, Q. Ding, H. Zheng, H. Xu, Z. Wei, *et al.*, "Uni-Mol: A Universal 3D Molecular Representation Learning Framework," *ChemRxiv*, 2022, doi:10.26434/chemrxiv-2022-jjm0j.
- [3] B. Yu, F. N. Baker, Z. Chen, X. Ning, and H. Sun, "LlaSMol: Advancing Large Language Models for Chemistry with a Large-Scale, Comprehensive, High-Quality Instruction Tuning Dataset," *arXiv preprint arXiv:2402.09391*, 2024. Available: <https://arxiv.org/abs/2402.09391>
- [4] S. Liu, J. Wang, Y. Yang, C. Wang, L. Liu, H. Guo, and C. Xiao, "ChatGPT-powered Conversational Drug Editing Using Retrieval and Domain Feedback," *arXiv preprint arXiv:2305.18090*, 2023. Available: <https://arxiv.org/abs/2305.18090>

```

def forward(self, smiles_tokens, text_tokens):
    # Encoder forward pass / Get SMILES embeddings
    smiles_tokens = {k: v.to('cuda') for k, v in smiles_tokens.items()}
    text_tokens = {k: v.to('cuda') for k, v in text_tokens.items()}

    mol_encoder_output = self.mol_encoder(**smiles_tokens)
    smiles_embedding = mol_encoder_output['last_hidden_state'][:, 0, :]
    smiles_projection = self.linear_project(smiles_embedding).unsqueeze(1)

    # Get embeddings from LLM for the question
    embedding_layer = self.llm_model.model.model.embed_tokens
    llm_embeddings = embedding_layer(text_tokens['input_ids']).squeeze(1)

    # Concatenate encoder and LLM embeddings
    combined_embeddings = torch.cat((smiles_projection, llm_embeddings), dim=1)

    attention_mask = torch.full((smiles_projection.shape[0], combined_embeddings.shape[1], combined_embeddings.shape[1]), float('inf'), device='cuda')
    attention_mask[:, 0, 0] = 0. # SMILES mask for itself
    for i in range(1, combined_embeddings.shape[1]):
        attention_mask[:, i, 0:i+1] = 0. # From SMILES to current token (inclusive)

    # Pass through Mistral's transformer Layers with LoRA adjustments
    output = self.llm_model(inputs_embeds=combined_embeddings, attention_mask=attention_mask)

    return output

```

Figure 22: Modified forward function which allows for the molecular token to be added.

- [5] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. Le Scao, T. Lavril, T. Wang, T. Lacroix, and W. El Sayed, “Mistral 7B,” *arXiv preprint arXiv:2310.06825*, 2023. Available: <https://arxiv.org/abs/2310.06825>
- [6] T. Dettmers, A. Pagnoni, A. Holtzman, and L. Zettlemoyer, “QLoRA: Efficient Finetuning of Quantized LLMs,” *arXiv preprint arXiv:2305.14314*, 2023. Available: <https://arxiv.org/abs/2305.14314>

## 14 MaSTeA: Materials Science Teaching Assistant

Authors: *Defne Ciri, Abhijeet S. Gangan, Mohd Zaki*

### 14.1 Dataset Details

We take 650 questions from materials science question answering dataset (MaScQA) [1], which required undergraduate-level understanding to solve them. The authors classified them into four types based on their structure: Multiple choice questions (MCQs), Match the following type questions (MATCH), Numerical questions where options are given (MCQN), and numerical questions (NUM): see Table 3. MCQs are generally conceptual, given four options, out of which mostly one is correct, and sometimes more than one option is also correct. In MATCH, two lists of entities are given, which are to be matched with each other. These questions are also provided with four options, out of which one has the correct set of matched entities. In MCQN, the question has four choices, out of which the correct one is identified after solving the numerical problem stated in the question. The NUM type questions have numerical answers, rounded to the nearest integer or floating-point number as specified in the questions.

To understand the performance of LLMs from a domain perspective, the questions were classified into 14 topical categories [1]. The database can be accessed at <https://github.com/M3RG-IITD/MaScQA>

### 14.2 Methodology

Our objective was to automate the evaluation of both open-source and proprietary LLMs on materials science questions from the MaScQA dataset and provide an interactive interface for students to solve these questions. We evaluated the performance of several language models, including LLAMA3-8B, HAIKU, SONNET, GPT-4, and OPUS, across the 14 various categories such as characterization, applications, properties, and behavior.

The evaluation involved:

- **Extracting corresponding values:** For multiple-choice questions, correct answer options were extracted using regular expressions to compare model predictions against the correct choices.
- **Prediction verification:** For numerical questions, the predicted value was checked against a specified range or exact value. For multiple-choice questions, the predicted answer was verified against the correct option or the extracted corresponding value.
- **Calculating accuracy:** Accuracy was calculated for each question type and topic, and the overall accuracy across all questions was computed.

The results of the evaluation are summarized in Table 4, which presents the accuracy of the various models for different question types and topics. The *opus* variant of Claude consistently outperformed the others, achieving the highest accuracy in most categories. GPT-4 also showed strong performance, particularly in topics related to material processing and fluid mechanics.

Our interactive web app, MaSTeA (Materials Science Teaching Assistant), developed using Streamlit, allows easy model testing to identify LLMs' strengths and weaknesses in different materials science subfields. The results suggest that there is significant room for improvement to enhance the accuracy of language models in answering scientific questions. Once these models become more reliable, MaSTeA could be a valuable tool for students to practice answering questions and learn the steps to get to the answer. By analyzing LLM performance, we aimed to guide future model development and pinpoint areas for improvement.

Our code and application can be found at:

- <https://github.com/abhijeetgangan/MaSTeA>
- <https://mastea-nhwpzz8fehvc9b3n5bhzya.streamlit.app/>

## References

[1] Zaki, M., & Krishnan, N. A. (2024). MaScQA: investigating materials science knowledge of large language models. *Digital Discovery*, 3(2), 313-327.

Table 3: Sample questions from each category: (a) multiple choice question (MCQ), (b) matching type question (MATCH), (c) numerical question with multiple choices (MCQN), and (d) numerical question (NUM). Correct answers are in bold.

<p>Floatation beneficiation is based on the principle of</p> <p><b>(A) Mineral surface hydrophobicity</b></p> <p>(B) Gravity difference</p> <p>(C) Chemical reactivity</p> <p>(D) Particle size difference</p>	<p>A peak in the X-ray diffraction pattern is observed at <math>2\theta = 78^\circ</math>, corresponding to <math>/311</math> planes of an fcc metal, when the incident beam has a wavelength of 0.154 nm. The lattice parameter of the metal is approximately</p> <p>(A) 0.6 nm</p> <p><b>(B) 0.4 nm</b></p> <p>(C) 0.3 nm</p> <p>(D) 0.2 nm</p>
<p>(a) Multiple choice question (MCQ)</p>	<p>(c) Numerical question with multiple choices (MCQN)</p>
<p>Match the composite in Column I with the most suitable application in Column II.</p> <p>Column I: (P) Glass fibre reinforced plastic, (Q) SiC particle reinforced Al alloy, (R) Carbon-carbon composite, (S) Metal fibre reinforced rubber:</p> <p>Column II: (1) Missile cone heads, (2) Commercial automobile chassis, (3) Airplane wheel tyres, (4) Car piston rings, (5) High performance skate boards:</p> <p>(A) P-4, Q-5, R-1, S-2 (B) P-3, Q-5, R-2, S-4</p> <p><b>(C) P-5, Q-4, R-1, S-3</b> (D) P-4, Q-2, R-3, S-1</p>	<p>The third peak in the X-ray diffraction pattern of a face-centered cubic crystal is at <math>2\theta</math> value of <math>45^\circ</math>, where <math>2\theta</math> is the angle between the incident and reflected rays. The wavelength of the monochromatic X-ray beam is <math>1.54 \text{ \AA}</math>. Considering first-order reflection, the lattice parameter (in <math>\text{\AA}</math>) of the crystal is? (Round off to two decimal places)</p> <p><b>Ans. 5.64 to 5.73</b></p>
<p>(b) Matching type question (MATCH)</p>	<p>(d) Numerical question (NUM)</p>

Table 4: Accuracy of Language Models by Topic

Topic	# Questions	LLaMA-3-8b	Haiku	Sonnet	OPUS	GPT4
Thermodynamics	114	37.72	47.37	55.26	<b>73.68</b>	57.02
Atomic structure	100	32	40	49	<b>64</b>	59
Mechanical behavior	96	22.92	41.67	52.08	<b>71.88</b>	43.75
Material manufacturing	91	43.96	57.14	56.04	<b>80.22</b>	68.13
Material applications	53	52.83	64.15	77.36	<b>92.45</b>	86.79
Phase transition	41	31.71	46.34	<b>65.85</b>	70.73	63.41
Electrical properties	36	33.33	25	55.56	<b>72.22</b>	44.44
Material processing	35	48.57	54.29	74.29	<b>88.57</b>	<b>88.57</b>
Transport phenomena	24	37.5	<b>70.83</b>	58.33	87.5	62.5
Magnetic properties	15	26.67	46.67	46.67	<b>66.67</b>	60
Material characterization	14	78.57	57.14	85.71	<b>92.86</b>	71.43
Fluid mechanics	14	21.43	50	57.14	78.57	<b>85.71</b>
Material testing	9	77.78	66.67	<b>100</b>	<b>100</b>	<b>100</b>
Miscellaneous	8	62.5	62.5	62.5	<b>75</b>	62.5

## 15 LLMY-Way

Authors: Ruijie Zhu, Faradawn Yang, Andrew Qin, Suraj Sudhakar, Jaehee Park, Victor Chen

### 15.1 Introduction

In the academic realm, researchers frequently present their work and that of others to colleagues and lab members. This task, while essential, is fraught with difficulties. For example, below are three challenges:

1. Reading and understanding research papers: Comprehending the intricacies of a research paper can be daunting, particularly for interdisciplinary subjects like materials science.
2. Creating presentation slides: Designing slides that effectively communicate the content requires significant effort, including remaking slides, sourcing images, and determining optimal text and image placement.
3. Tailoring to the audience: Deciding on the appropriate level of technical vocabulary and the number of slides needed to fit within a given time limit adds another layer of complexity.

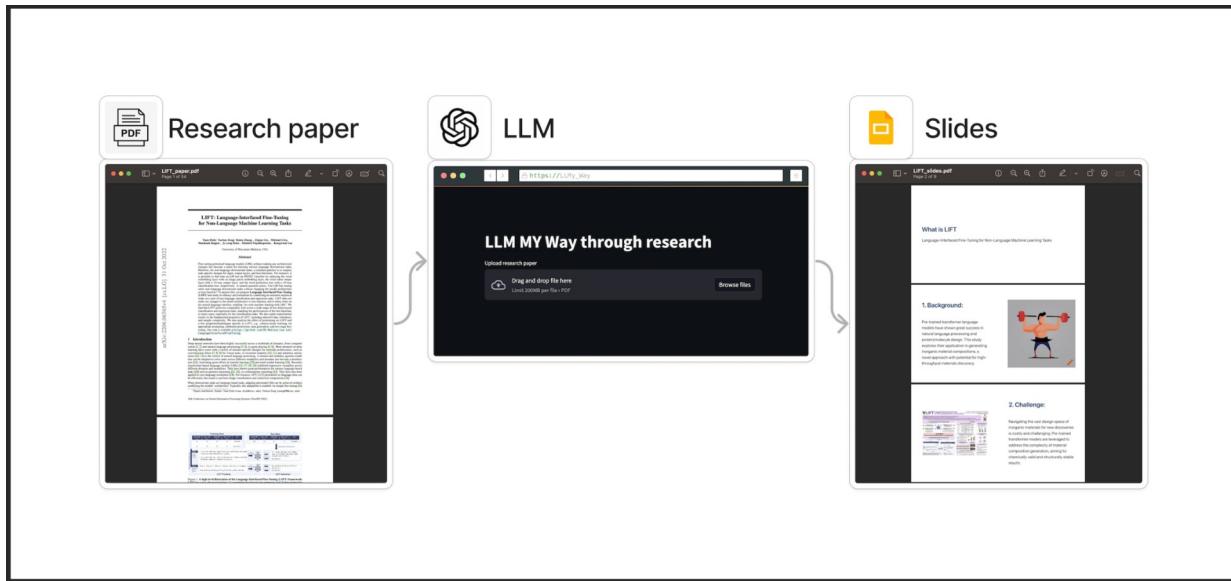


Figure 23

These challenges can be effectively addressed using large language models, which can streamline and automate the text summarization and slide creation process. LLMY Way leverages the power of GPT-3.5-turbo to automate the creation of academic slides from research articles. The methodology involves primarily three steps:

### 15.2 Structured Summarization

We prompt the large language model to generate a structured summary of the research paper, adhering to the typical sections of an academic paper: background, challenge, current status, method, result, conclusion, and outlook. Each section is summarized in under a certain number of words to ensure brevity and clarity.

Example Prompt: Read the paper, and summarize in `num_words` words each about the following:  
1. Background  
2. Challenge  
3. Current status  
4. Method  
5. Result  
6. Conclusion  
7. Outlook

### 15.3 Slide Generation

To create slides, we format the language model's output in a specific manner, using symbols to denote slide breaks. This output is then parsed and converted into a Markdown file, where additional images and text formatting are applied as needed. The formatted Markdown file is subsequently transformed into PDF slides using Marp. Example output formatting:

```
# Background
Summary of the background here.
---
# Challenge
Summary of the challenge here.
---
```

### 15.4 Customization for Audience and Time Limit

LLMy Way allows customization based on the target audience's expertise level (expert or non-expert) and the presentation time limit (e.g., 10 minutes, 15 minutes). This information is incorporated into the initial generation phase, ensuring that the content and slide count are appropriately tailored. This feature is to be implemented.

### 15.5 Conclusion

LLMy Way represents a significant step forward towards the automation of academic presentation preparation. By leveraging the structured nature of scientific papers and the capabilities of advanced large language models, our tool addresses common pain points faced by researchers. The current implementation of our framework can be summarized into three consecutive steps. First, the research paper is parsed by LLM, which is summarized into predefined sections. Next, the summarized texts are converted into Markdown. Finally, Marp is used to generate the final slide deck in the PDF format. The current implementation of our framework uses GPT-3.5-turbo, but it can be adapted to other language models as needed. We also support the output format of LaTeX to fit the needs of many researchers. Future work will focus on further refining the tool, incorporating user feedback, and exploring additional customization options.

## References

- [1] OpenAI. (2024). GPT-3.5-turbo. <https://openai.com/api/>
- [2] Marp. (2024). Markdown Presentation Ecosystem. <https://marp.app/>

# 16 WaterLLM: Creating a Custom ChatGPT for Water Purification Using Prompt-Engineering Techniques

Authors: Viktoriia Baibakova, Maryam G. Fard, Teslim Olayiwola, Olga Taran

## 16.1 Introduction

Drinking water pollution is a growing environmental problem that, ironically, comes from an increase in industrial production of new materials and chemicals. Common pollutants include heavy metals, perfluorinated compounds, microplastic particles, excreted medicinal drugs from hospitals, agricultural runoff and many others [1]. The communities that suffer the most from water contamination often lack the plumbing infrastructure necessary for centralized water analysis and treatment. Decentralized and localized water treatments, based on resources available to the communities, can alleviate the problem. Since the resources can vary greatly, from well equipped analytical facilities to low-cost DIY solutions, a knowledge base that can rapidly provide information relevant for specific situations is needed. Here we show a prototype Large Language Model (LLM) chatbot that can take a variety of inputs about possible contaminants (ranging from detailed LC/MS analysis to general description of the situation) and propose the best solution to the water treatment for the particular case based on contaminant composition, cost and resources availability. We employed a Retrieval Augmented Generation (RAG) capability of ChatGPT to answer questions about possible water treatments based on the latest scientific literature.

For this project, we focused on advanced oxidation procedures (AOPs) for water purification from microplastics (MPs). In recent times, MPs have received significant attention globally due to their widespread presence in various species' bodies, environmental media, and even bottled drinking water, as frequently documented [2]. Numerous trials have been conducted and reported on the use of AOPs for breaking down diverse persistent microplastics as wastewater treatment methodologies. However, there remains a lack of guidelines on selecting the most suitable and cost effective treatment method based on the characteristics of the contaminant, maximum removal percentage of MP.

The complexity of existing research on AOPs for MPs can be tackled with LLMs enhanced with RAG. RAG allows to augment LLM's knowledge and achieve state of the art results on knowledge-intensive tasks. One straightforward modern way to implement LLM with RAG is through configuring a custom chatGPT. We uploaded current scientific papers under the "Knowledge" for RAG and tailored chatbot performance with prompt-engineering techniques such as grounding, context, and chain-of-thought reasoning to ensure that it delivers accurate, detailed, and useful information.

## 16.2 Grounding

To make sure that chatbot provides accurate and scientifically valid answers, we loaded the latest research on microplastic pollution remediation for RAG and implemented grounding in chat prompt. To collect the data, we gathered the initial set of 10 review articles from the expert in the field that talk about water purification using Advanced Oxidation Process. Then, we manually found 112 scientific articles discussing the specific treatment procedure. This way, our dataset has studies on different pollution sources like laundry, hospitals, industry, different pollutant types and their descriptive characteristics like size, shape, color, and different treatment methods like Ozonation, Fenton Reaction, UV, Heat-Activated Persulfate. We merged all papers into 8 pdfs to meet chatGPT "Knowledge" restrictions in files number, size and length. With grounding, we aim to anchor the chatbot's responses in concrete and factual knowledge. We explicitly asked chatGPT to avoid giving wordy broad generalized answers and to provide concise scientific details using the files uploaded under Knowledge.

## 16.3 Context

We provided context for the chatbot to understand and respond appropriately to user queries. We specified that Chatbot has expert-level knowledge on MPs and water purification strategies from MPs and other contaminants. We defined a user as a technician with basic knowledge on chemical engineering that needs to choose and apply a purification method. We set that the communication between Chatbot and User should

be in the form of interactive dialog. It means that Chatbot should ask follow-up questions from the user and should finally return an accurate purification protocol with all the details that can be reproduced in the experiment.

## 16.4 Chain-of-Thought Reasoning

To further ensure that the chatbot considers all necessary aspects before providing a solution, we employed the chain-of-thought reasoning by breaking down the problem-solving process into sequential steps. Get the source of contamination. Ask what the pollutant particle is and suggest evaluation tests if it is unknown. Ask if the characteristics of pollutants are known: size, shape, type. Suggest the most effective purification approach that will get rid of the largest percentage of the pollutants and estimate the price. Adjust if it is too expensive. Inquiry about the post-treatment analysis. If unsatisfactory, go to the previous step. End of conversation: Chatbot should provide a table with all used purification methods and their parameters for established contaminants.

These techniques allowed to boost the custom GPT performance, and WaterLLM demonstrated performance approved by the expert in the field.

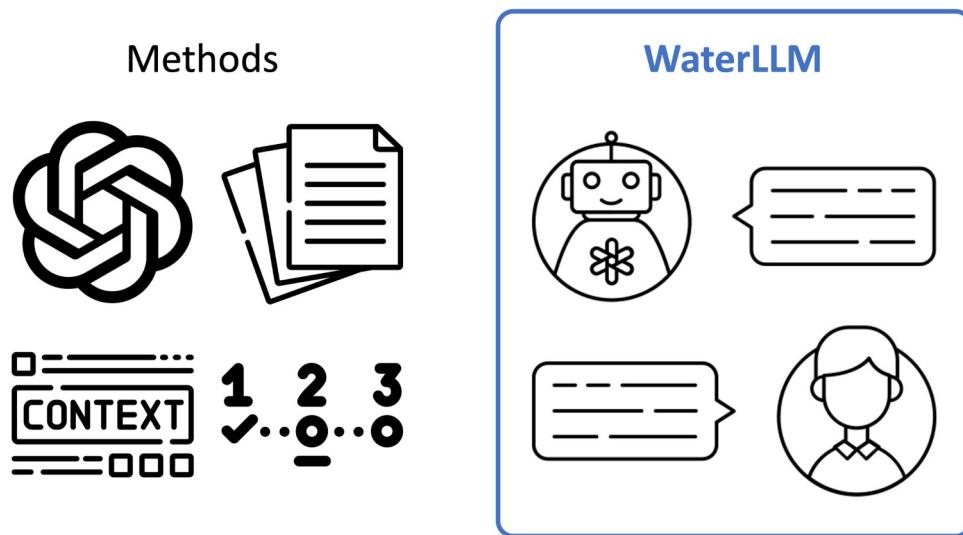


Figure 24: WaterLLM approach: custom chatGPT with RAG from scientific papers, context and chain-of-thought allowed for interactive dialog with the user anchored to science.

## References

- [1] Mishra, R. K.; Mentha, S. S.; Misra, Y.; Dwivedi, N. Emerging Pollutants of Severe Environmental Concern in Water and Wastewater: A Comprehensive Review on Current Developments and Future Research. *Water-Energy Nexus* 2023, 6, 74–95. <https://doi.org/10.1016/j.wen.2023.08.002>
- [2] Li, Y.; Peng, L.; Fu, J.; Dai, X.; Wang, G. A Microscopic Survey on Microplastics in Beverages: The Case of Beer, Mineral Water and Tea. *Analyst* 2022, 147 (6), 1099–1105. <https://doi.org/10.1039/D2AN00083K>

Hi! I have a dirty water. What to do?

To help you clean the dirty water, I'll need more specific information. Let's go through the details step-by-step to identify the best treatment method:

1. Source of Contamination:
  - What is the source of the dirty water? (e.g., laundry runoff, industrial discharge, hospital wastewater, etc.)
2. Pollutant Particles:
  - Do you know the specific pollutants in the water? If not, can you describe the general contaminants (e.g., visible particles, chemical odor, oily texture)?
3. Characteristics of Pollutants:
  - Do you have details on the size, shape, and type of the pollutants? If not, we can suggest some tests to determine these characteristics.

industrial discharge

Great, we are dealing with industrial discharge. Let's proceed with the next steps:

I don't know

Since the specific pollutants in the industrial discharge are unknown, let's start by identifying the types of contaminants. Here are some evaluation tests you can perform:

1. Visual Inspection:
 

I have an access to the lab, can you suggest more precise tests?

Since you have access to a lab, here are some more precise tests to identify the contaminants in the industrial discharge:

1. Gravimetric Analysis:
  - Total Suspended Solids (TSS): Filter a known volume of water through a pre-weighed filter, dry the filter, and weigh it again to determine the amount of suspended solids.
  - Total Dissolved Solids (TDS): Evaporate a known volume of filtered water and weigh the residue to determine dissolved solids.

We found fragments, fibers and films. We looked at the microscope and they all are 0.1-1 mm in size. Most of them are white. Can you infer what kind of plastic can it be?

Based on the description of the fragments, fibers, and films being 0.1-1 mm in size and mostly white, it is likely that they are common types of plastics used in industrial processes. Here are some possible types of plastics that match your description:

1. Polyethylene (PE):
  - **High-Density Polyethylene (HDPE):** Often used in containers, pipes, and industrial applications. It can appear as white or translucent fragments or films.
  - **Low-Density Polyethylene (LDPE):** Commonly used in plastic bags, film wrap, and some industrial liners. It is flexible and can also be white.
2. Polypropylene (PP):
  - **PE:** Burns with a blue flame with a yellow tip, drips, and smells like paraffin.
  - **PP:** Burns with a blue flame with a yellow tip, drips, and smells like burning oil.
  - **PS:** Burns with a sooty flame, does not drip, and has a sweet smell.
  - **PVC:** Burns with a green flame, emits hydrogen chloride gas (pungent smell), and self-extinguishes.

We did a burn test, and it burns with a blue flame with a yellow tip, drips, and smells like paraffin

Based on the burn test results—burning with a blue flame with a yellow tip, dripping, and smelling like paraffin—it is most likely that the plastic contaminants are polyethylene (PE).

Step	Details
Contaminants	Microplastics (100–300 microns, blue, black, yellow)
Treatment Method	Combined Fenton and UV Treatment
Reagents	H <sub>2</sub> O <sub>2</sub> (10 mL), FeSO <sub>4</sub> ·7H <sub>2</sub> O (0.1 g), H <sub>2</sub> SO <sub>4</sub> (1 mL), NaOH (1 g)
Equipment	UV Lamp (254 nm), Quartz Vessel, Magnetic Stirrer, Filter Paper
Reaction Time	60-120 minutes
Total Cost	\$2.74 per treatment
Outcome	Effective reduction of microplastics

Figure 25: Sample of the WaterLLM communication with the User.

## 17 yeLLowhaMMer: A Multi-modal Tool-calling Agent for Accelerated Research Data Management

*Authors:* Matthew L. Evans, Benjamin Charmes, Vraj Patel, Joshua D. Bocarsly

As scientific data continues to grow in volume and complexity, there is a great need for tools that can simplify the job of managing this data to draw insights, increase reproducibility, and accelerate discovery. Digital systems of record, such as electronic lab notebooks (ELN) or laboratory information management systems (LIMS), have been a great advancement in this area. However, complex tasks are still often too laborious, or simply impossible, to accomplish using graphical user interfaces alone, and any barriers to streamlined data management often lead to lapses in data recording.

As developers of the open-source *datalab* [1] ELN/LIMS, we explored how large language models (LLMs) can be used to simplify and accelerate data handling tasks in order to generate new insights, improve reproducibility, and save time for researchers. Previously, we made progress toward this goal by developing a conversational assistant, named Whinchat [2], that allows users to ask questions about their data. However, this assistant was unable to take action with a user’s data. Here, we developed yeLLowhaMMer, a multimodal large language model (MLLM)-based data management agent capable of taking free-form text and image instructions from users and executing a variety of complex scientific data management tasks.

Our agent is powered by a low-cost commercial MLLM (Anthropic’s Claude 3 Haiku) used within a custom agentic infrastructure that allows it to write and execute Python code that interacts with *datalab* instances via the *datalab-app* package. In typical usage, a yeLLowhaMMer user might instruct the agent: “Pull up my 10 most recent sample entries and summarize the synthetic approaches used.” In this case, the agent will attempt to write and execute Python code using the *datalab* API to query for the user’s samples in the *datalab* instance and write a human-readable summary. If the code it generates gives an error (or does not give sufficient information), the agent can iteratively rewrite the program until the task is accomplished successfully.

Furthermore, we leveraged the powerful multimodal capabilities of the latest MLLMs to allow for prompts that include visual cues. For example, a user may upload an image of a handwritten lab notebook page and ask that a new sample entry be added to the *datalab* instance. The agent uses its multimodal capabilities to “read” the lab notebook page (even if it is a messy/unstructured page), adds structure to the information it finds by massaging it into the form requested by the *datalab* JSON schema, then writes a Python snippet to ingest the new sample into the *datalab* instance. Notably, we found that even the inexpensive, fast model we used (Claude 3 Haiku) was able to perform sufficiently well at this task, while larger models may be explored in the future to allow for more advanced applications (though with slower speed and greater cost). We believe the capabilities demonstrated by yeLLowhaMMer show that MLLM agents have the potential to greatly lower the barrier to advanced data handling in experimental materials and chemistry laboratories. This proof-of-concept work is accessible on GitHub at [bocarsly-group/llm-hackathon-2024](https://github.com/bocarsly-group/llm-hackathon-2024), with ongoing work at [datalab-org/yellowhammer](https://github.com/datalab-org/yellowhammer).

yeLLowhaMMer was built upon several open-source software packages. The *codebox-api* Python package was used to set up a local code sandbox that the agent has access to in order to safely read and save files, install Python packages, and run the code generated by the model. The *datalab-app* Python package was used to interact with *datalab* instances. An MLLM-compatible tool was designed to allow the model to use function-calling capabilities, write, and execute code within the sandbox. LangChain was used as a framework to interact with the commercial MLLM APIs and build the agentic loop. Streamlit was used to build a responsive GUI to show the conversation and upload/download the files from the codebox. A customized Streamlit callback was written to display the code and files generated by the agent in a user-friendly manner.

An interesting challenge in the development of yeLLowhaMMer was the creation of a system prompt that would enable the agent to reliably generate robust code using the *datalab-app* package, which is a recent library not included in the training of the commercial models at the time of writing. Initially, we copied the existing documentation for the *datalab-app* into the system prompt, but we found that the code generated by the model was not working very well. Instead, it was helpful to produce a simplified version of the documentation that removed extraneous information and gave a few concrete examples of scripts. Additionally, we provided an abridged version of the *datalab* schemas in the JSON Schema format in the system prompt, which was necessary for the generation of compliant data to be inserted into *datalab*.

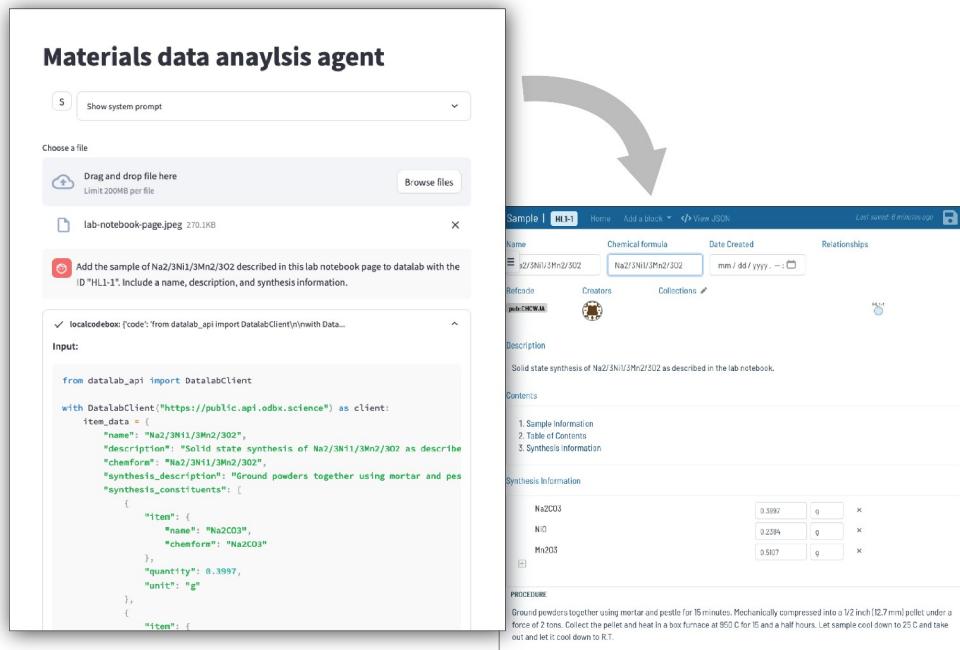


Figure 26: The yeLLowhaMmer multimodal agent can be used for a variety of data management tasks. Here, it is shown automatically adding an entry into the *databar* lab data management system based on an image of a handwritten lab notebook page.

Overall, the yeLLowhaMmer system prompt amounts to around 12,000 characters (corresponding to about 3200 tokens using Claude’s tokenizer). Given the large context windows of the current generation of MLLMs (e.g., 200k tokens for Claude 3 Haiku), this size of prompt is feasible for fast, extended conversations involving text, generated code, and images. In the future, we envision that library maintainers may wish to increase the utility of their libraries by maintaining two parallel sets of documentation: the standard human-readable documentation, and an abridged `agents.txt` for LLMs (<https://llmstxt.org>) file that can be used by ML agents to write high-quality code using that library.

Going forward, we will undertake further prototyping to incorporate MLLM-based agents more tightly into our data management workflows, and to ensure that data curated or modified by such an agent will be appropriately ‘credited’ by, for example, visually demarcating AI-generated content, and providing UI pathways to verify or ‘relabel’ such data in an efficient manner. Finally, we emphasize the great progress made within the last year in MLLMs themselves, which are now able to handle audio and video content in addition to text and images. These will allow MLLM agents to use audiovisual data in real-time to provide new user interfaces. Based on these promising developments, we believe that data management platforms are well-placed to help bridge the divide from physical to digital data recording.

## References

- [1] M. L. Evans and J. D. Bocarsly. *databar*, July 2024. URL <https://github.com/databar-org> doi:10.5281/zenodo.12545475.
- [2] Jablonka et al. 14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon. *Digital Discovery*, 2023. doi:10.1039/D3DD00113J.

## 18 LLMads

*Authors: Sarthak Kapoor, Jose M. Pizarro, Ahmed Ilyas, Alvin N. Ladines, Vikrant Chaudhary*

Parsing raw data into a structured (meta)data standard or schema is a major Research Data Management (RDM) topic. While defining F.A.I.R. (Findable, Accessible, Interoperable, and Reusable) metadata schemas are the key to RDM, these empty fields must be populated. This is typically done in two ways:

- Fill a schema manually using electronic or physical lab notebooks, or
- Create scripts that read the input/output raw files and parse them into the data schema.

The first option is used in a lab setting where data is entered in real-time as it is generated. The second option is used when data files are available, albeit in an incompatible format to fill the schema directly. These can be measurement files coming from instruments or files generated from simulations. Specific parsers for each raw file type can transfer large amounts of data into schemas, making them essential for automation and big-data management. However, implementing parsers for all the possible raw files to a fill schema can be laborious and time-consuming. It requires expert knowledge of the structure of the raw files and regular maintenance to keep up with new versions of raw files. In this work, we attempted to substitute parsers with Large Language Models (LLMs).

We investigated whether LLMs can be used to parse data into structured schemas, thus relieving the need for coding parsers. As an example, we used raw files from X-ray diffraction (XRD) measurements from three different instrument vendors (Bruker, Rigaku, and Panalytical). We defined a Pydantic model for our structured data schema and used the pre-trained Mixtral-8x7b from Groq. The data schemas are provided to the LLM using the function-calling mechanism. The schema is constructed by defining a Pydantic base model class and their fields or attributes with well-defined types and descriptions. The LLM tries to extract data for each variable from the raw files that matches these descriptions. Considering the size of the raw files and the token limitation of LLMs, we decided to create the following workflow:

- Break the raw file contents into chunks.
- Prompt the LLM with the initial chunk.
- Generate a response from the LLM and populate the schema.
- Prompt the LLM with the next chunk along with the previous response.

We found that when populating the schema, the LLM was correctly extracting the values in cases where the data types were `float` and `str`. This was the case for the `XRDSettings` class. However, when parsing values with a data type `list[float]`, the LLM was often unable to extract the data in the expected format. This occurred when populating `XRDResult` class with the intensities data. The LLM output included non-numeric characters like `\n` or `\t`, along with hallucinated data values. By providing the previous response along with the new chunk of data in the prompts, we incorporated some degree of context. We found that using smaller chunk sizes led to the rapid replacement of the populated data. Sometimes, the correct data was replaced by hallucinated values.

We used LangChain to build our models and prompt generators. The code is openly available on Github: <https://github.com/ka-sarthak/llmads>. Our work uses prompt engineering and function-calling. Future work into tuning the model temperature and fine-tuning could be explored to combat hallucination. Our work also indicates a need for human intervention to verify if the schema was filled correctly and to correct it when necessary. Nevertheless, our prompting strategy proves to be a valuable tool as it manages to initialize the schema properly for non-vectorial fields, all at the minimal effort of providing a structured schema and the raw files.

## 19 NOMAD Query Reporter: Automating Research Data Narratives

**Authors:** *Nathan Daelman, Fabian Schöppach, Carla Terboven, Sascha Klawohn, Bernadette Mohr*

Materials science research data management (RDM) platforms and structured data repositories contain large numbers of entries, each composed of property-value pairs. Users query these repositories by specifying property-value pairs to find entries matching specific criteria. While this guarantees that all returned entries have at least the queried properties, they do not provide context or insights into the structure and variety of other data present in them.

Traditionally, it is up to the data scientist to examine the returned properties and interpret the overall response. To assist with this task, we use a LLM to create context-aware reports based on the properties and their meanings. We build and tested our “Query Reporter” [1] prototype on the NOMAD [2] repository, which stores heterogeneous materials science data, including metadata on scientific methodologies, atomistic structures, and materials properties.

We developed the NOMAD Query Reporter [1] as a proof-of-concept. It fetches and analyzes entries and produces a summary of the used methodology and standout results. It does so in a scientific style, lending itself as the basis for a “methods” section in a journal article. Our agent uses a retrieval-augmented generation (RAG) approach [3], which enriches an LLM’s knowledge of external DB data without performing retraining or fine-tuning. To allow its safe application on private and unpublished data, we use a self-hosted Ollama instance running Meta’s Llama3 70B [4] model.

We tested the agent on publicly available data from NOMAD. To manage Llama3’s context window of 8,000 tokens, the entries are collected as rows into a Pandas dataframe. Each row (e.g., entry) is individually passed on to the LLM via the chat-completion API. Instead of a single message, it accepts a multi-turn conversation that simulates several demarcated roles. We use the “system” and “user” roles of the chat to reinforce the retention of parts of the previous summary. This approach generally conforms to the Naive RAG category in Gao et al.’s classification [3]. For a step-by-step overview, see Figure 27.

We used two kinds of data for testing: (a) homogeneous, property-value pairs of computational data; and (b) heterogeneous text typed properties of solar cell experiments, often formatted as in-text tables. We engineered different prompts for each kind. The agent performed better on the homogeneous than the heterogeneous data. Here, summaries would often suffer from irrelevant threads, or even hallucinations. We theorize that homogeneous data maps more consistently onto our predefined dataframe columns, which aids the LLM in interpreting follow-up messages. Still, we could not improve the performance for heterogeneous data within the hackathon.

In short, the NOMAD Query Reporter demonstrates that the combined approach of filtering and RAG can effectively summarize collections of limited-size, structured data directly stored in research data repositories, allowing for automated drafting of methods and setups for publications at a consistent level of quality and style. These results suggest applicability for other well-defined materials science APIs, such as the OPTIMADE standard [5]. Follow-up work includes investigating the impact of Advanced RAG strategies [3].

### 19.1 Acknowledgements

N. D., S. K., and B. M. are members of the NFDI consortium FAIRmat, funded by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) in project 460197019. F. S. acknowledges funding received from the SolMates project, which has been supported by the European Union’s Horizon Europe research and innovation program under grant agreement No 101122288C. T. is supported by the German Federal Ministry of Education and Research (BMBF, Bundesministerium für Bildung und Forschung) in the framework of the project Catlab 03EW0015A.

## References

[1] NOMAD: A distributed web-based platform for managing materials science research data. M. Scheidgen, et al., JOSS, 8(90), 5388 (2024), [doi.org/10.21105/joss.05388](https://doi.org/10.21105/joss.05388)

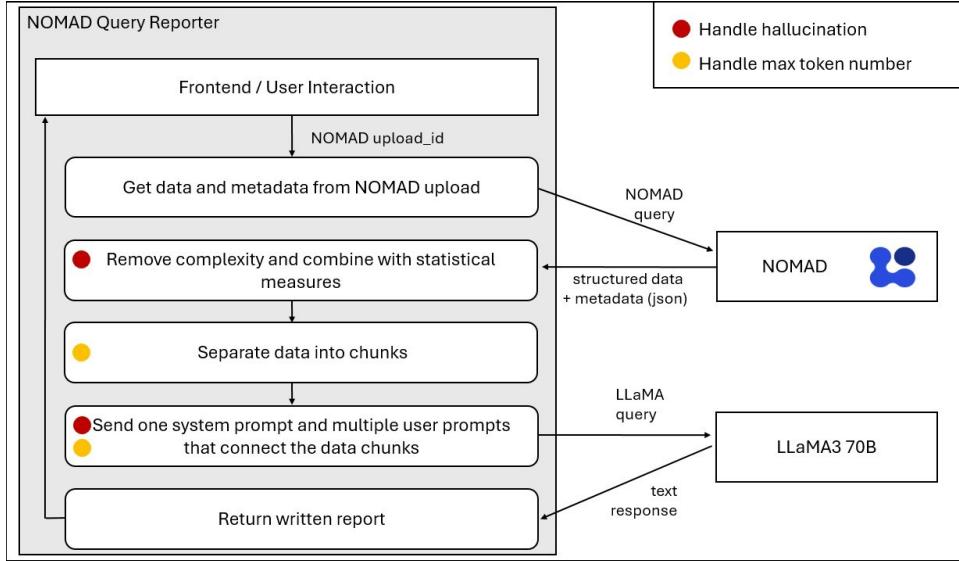


Figure 27: Flowchart of the Query Reporter usage, including the back-end interaction with external resources, i.e., NOMAD and Llama. Intermediate steps managing hallucinations or token limits are marked in red and orange, respectively.

[2] [https://github.com/ndaelman-hu/nomad\\_query\\_reporter](https://github.com/ndaelman-hu/nomad_query_reporter)

[3] Retrieval-Augmented Generation for Large Language Models: A Survey. Gao, Y., et al., arXiv (2024), [doi.org/10.48550/arXiv.2312.10997](https://doi.org/10.48550/arXiv.2312.10997)

[4] Llama: open and efficient foundation language models. H. Touvron, et al., arXiv (2023), [doi.org/10.48550/arXiv.2302.13971](https://doi.org/10.48550/arXiv.2302.13971)

[5] Development and applications of the OPTIMADE API for materials discovery, design, and data exchange. Evans, M. L., et al., Digital Discovery (2024), DOI: [doi.org/10.1039/D4DD00039K](https://doi.org/10.1039/D4DD00039K)

# 20 Speech-schema-filling@creating Structured Data Directly from Speech

Authors: *Hampus N\"asstr\"om, Julia Schumann, Michael*

*G\"otte, Jos\'e A. M\'arquez*

## 20.1 Introduction

As the amount of materials science data being created increases, so do the efforts to make this data Findable, Accessible, Reusable, and Interoperable (FAIR) [1]. One pragmatic approach to creating FAIR data is by defining so-called data schemas for the various types of data being recorded. These schemas can then be used in both input tools like electronic lab notebooks (ELNs) and storage solutions like data repositories to create structured data. One widely adopted standard for writing data schemas is the so-called JSON Schema [2]. JSON Schema allows us to define objects, such as, for example, a solution preparation experiment in the lab, with properties such as temperature and a list of solutes and solvents (see Figure 28a). These schemas can then be used to create forms in an ELN like NOMAD [3] (see Figure 28b). However, in a lot of lab situations, such as when working inside a glovebox, it is difficult to i) navigate the ELN and select the right form and ii) actually fill in the form with experimental data. In our experience, this usually results in users having to record their data later from memory or even in data not being recorded at all.

We propose a solution for this using LLMs to:

- Converting spoken language in the lab into text using advanced speech recognition technologies, such as OpenAI's Whisper.
- Based on the text select and fill the appropriate schema, enabling accurate data capture without manual text entry.

## 20.2 Speech recognition

The first step of converting speech into structured data is to record the speech and transcribe it into text. There are multiple options for recording audio, and the best solution depends on both the hardware and the operating system used. We use the `Recognize` class from the Python package `SpeechRecognition`[4] to perform the recording only, while leaving the actual speech recognition to OpenAI's Whisper. There is an open version of this available for download through the Python package `openai-whisper`[5].

## 20.3 Text to structured data

Once the speech has been converted to text we need to use this text to i) select the appropriate schema and ii) fill the schema. For this, we make use of a common feature in LLMs called "function calling", "API calling", or "tool use". This has been developed to allow LLMs to make valid API calls and, since OpenAPI is validated by it, uses JSON Schema to define the possible "functions", or "tools", that can be outputted. Since we use JSON Schema to define our data schemas we can simply supply these as the tools for the LLM to use. One way to do this in Python is to write a Pydantic schema for each of the data schemas and then convert this to a JSON Schema. For the Llama model we used, the python package `langchain-experimental`[6] has an `OllamaFunctions` class that can be imported from the `llms.ollama` `functions` sub-package. This class can then be used to instantiate the model and add the valid data schemas. Here is an example using LangChain and Llama3:

```
from langchain.experimental.llms.ollama import OllamaFunctions
model = OllamaFunctions(model="Llama3:70b", baseurl='...', format='json')
model = model.bind_toools(
    tools=[{
        "name": "solution_preparation",
        "description": "Schema for solution preparation",
        "parameters": SolutionPreparation.schema(),
    }]
)
```

```

        },
        {
            "name": "powderscaling",
            "description": "Schema for powder scaling",
            "parameters": Scaling.schema(),
        }
    ],
)

```

Where `SolutionPreparation` and `Scaling` are Pydantic models for our desired data schemas. Finally, this can be chained together with a prompt template and used to process the transcribed audio from before:

```

prompt = PromptTemplate.from_template(...)
chain_with_tools = prompt / model
response = chain_with_tools.invoke(transcribed_audio)

```

For `LangChain` the selected schema can be retrieved from:

```
schema = response.additional_kwargs['functional']['name']
```

And the filled instance from:

```
instance = json.loads(
    response.additional_kwargs['functional']['arguments'])
```

For `LangChain` the selected schema can be retrieved from:

```
schema = response.additional_kwargs['functional']['name']
```

And the filled instance from:

```
instance = json.loads(
    response.additional_kwargs['functional']['arguments'])
```

A detailed example of the text-to-structured data can be found as an iPython notebook on [github.com/hampusnasstrom/speech-schema-filter](https://github.com/hampusnasstrom/speech-schema-filter) together with an implementation of the audio recording and transcribing using Whisper. In conclusion, we believe that LLMs can be useful in labs where traditional ELNs are hard to operate by transcribing speech, selecting appropriate schemas, and filling the schemas to ultimately create structured data directly from speech.

## 20.4 Acknowledgements

H.N., J. S., and J. A. M. are part of the NFDI consortium FAIRmat funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project 460197019.

## References

- [1] Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016) <https://doi.org/10.1038/sdata.2016.18>
- [2] <https://json-schema.org/>
- [3] Scheidgen et al., (2023). NOMAD: A distributed web-based platform for managing materials science research data. *Journal of Open Source Software*, 8(90), 5388 <https://doi.org/10.21105/joss.05388>
- [4] <https://pypi.org/project/SpeechRecognition/>
- [5] <https://pypi.org/project/openai-whisper/>
- [6] <https://pypi.org/project/langchain-experimental/>

a)

```
{  
  "title": "SolutionPreparation",  
  "type": "object",  
  "properties": {  
    "temperature": {  
      "title": "Temperature",  
      "description": "The solution creation temperature",  
      "type": "number"  
    },  
    "solutes": {  
      "title": "Solutes",  
      "description": "The solutes used in the solution",  
      "type": "array",  
      "items": {  
        "type": "string"  
      }  
    },  
    ...  
  }  
}
```

b)

SolutionPreparation

temperature	<input checked="" type="checkbox"/> Unit °C
atmosphere	
method	
time	<input checked="" type="checkbox"/> Unit s
solutes	
solutes	
solute masses	
solute masses	<input checked="" type="checkbox"/> Unit mg
solvents	
solvents	
solvent volumes	
solvent volumes	<input checked="" type="checkbox"/> Unit ml

Figure 28: a) Part of a JSON Schema defining a data structure for a solution preparation. b) The schema converted to an ELN form in NOMAD [3].

# 21 Leveraging LLMs for Bayesian Temporal Evaluation of Scientific Hypotheses

Authors: Marcus Schwarting

## 21.1 Introduction

Science is predicated on empiricism, and a scientist uses the tools at their disposal to gather observations that support the veracity of their claims. When one cannot gather firsthand evidence for a claim, they must rely on their own assessment of evidence presented by others. However, the scientific literature on claim is often large and dense (particularly for those without domain expertise), and the scientific consensus on the veracity of a claim may drift over time. In this work we consider how large language models (LLMs), in conjunction with temporal Bayesian statistics, can rapidly provide a more holistic view of a scientific inquiry. We demonstrate our approach on the hypothesis that the material LK-99, which went viral after its discovery in April 2023, is in fact a room-temperature superconductor.

## 21.2 Background

Scientific progress requires a researcher to iteratively update their prior understanding based on new observations. The process of updating a statistical prior based on new information is the backbone of Bayesian statistics and is routinely used in scientific workflows [1]. Under such a model, a hypothesis has an inferred probability  $P_H \in (0, 1)$  of being true and will never be completely discarded ( $P_H = 0$ ) or accepted ( $P_H = 1$ ) but may draw arbitrarily close to these extrema. This inferred probability from a dataset is also a feature of assessing the power of a claim using statistical hypothesis tests [2], which are commonly used across most scientific disciplines.

Modelling the veracity of a scientific claim as a probability also stems from the work of philosopher Karl Popper [3]. Popper posits that a scientific claim should be falsifiable, such that it can be refuted by empirical evidence. If a scientific claim cannot be refuted, Popper argues that it is not proven, but gains further credibility from the scientific community. Scientific progress is not made by proving claims, but instead by hewing away false claims through observation until only those that withstand repeated scrutiny remain.

The history of science is littered with discarded hypotheses. Some of these claims were believed to be true for centuries before being dismissed (including geocentrism, phrenology, energeticism, and spontaneous generation). In this work, we focus on a claim by Lee, Kim, and Kwon that they had successfully synthesized a room-temperature superconductor, which they called LK-99 [4]. Such a material would necessitate altering the existing theory of superconductors at a fundamental level and would enable innovations that are currently beyond reach. LK-99 went viral in summer 2023 [5], but replication efforts quickly ran into issues [6]. Since the initial claim was made in April 2023, roughly 160 works have been published, and the scientific consensus now appears established: LK-99 is not a room-temperature superconductor.

## 21.3 Methods

Our dataset consists of the 160 papers on Google Scholar, ordered by publication date, that are deemed relevant to the hypothesis “LK-99 is a room-temperature superconductor.” For each paper abstract in our dataset, we construct an LLM prompt as follows to perform a zero-shot natural language inference operation [7]:

Given a Hypothesis and an Abstract, determine whether the Hypothesis is an ‘entailed’, ‘neutral’, or ‘contradicted’ by the Abstract. \n Hypothesis: /claim/ \n Abstract: /abstract/

We then wrote a regular expression to check the LLM response to make an assertion for each publication: “entailment,” “neutrality,” or “contradiction.” We use Llama2 to make these assertions on all 160 papers, which complete in under five minutes (on a desktop with an Nvidia GTX-1070 GPU).

Next, we constructed a temporal Bayesian model where, starting from an initial Gaussian prior  $N(\mu_P, \sigma_P^2)$  that models the likelihood of accepting the hypothesis, we can update with a Gaussian likelihood  $N(\mu_L, \sigma_L^2)$ .

Our likelihood is a Gaussian designed so that for a given publication, “entailment” acceptance probability is higher, “contradiction” pushes the acceptance probability lower, and “neutrality” leaves the acceptance probability the same. Our likelihood probability is further weighted according to the impact factor of the journal, with an impact factor floor set to accommodate publications that have not passed peer review. Retroactively, we could also weight by the number of citations, however we omit this feature in our analysis since this is inherently a post-hoc metric and would violate our temporal assessment. We update our Gaussian prior using the equations [8]:

$$\mu_P \leftarrow \frac{1}{\sigma_P^2} + \frac{1}{\sigma_L^2} \quad -1 \quad \frac{\mu_P}{\sigma_P^2} + \frac{\mu_L}{\sigma_L^2} ; \quad \sigma_P^2 \leftarrow \frac{1}{\sigma_P^2} + \frac{1}{\sigma_L^2} \quad -1$$

We are also able to specify the initial probability associated with the hypothesis ( $\mu_P$ ), as well as how flexible we are in changing our perspective based on new evidence ( $\sigma_P^2$ ). We select two initial probabilities: 50% and 20%, and fix standard deviations  $\sigma_P^2$  and  $\sigma_L^2$ . Assuming either “contradiction” or “entailment”, the update due to  $\mu_L$  then scales linearly with the publication impact factor. Finally, we can compare our temporal probability assessment with probabilities provided by the betting platform Manifold Markets [9], where players bet on the outcome of a successful replication of the LK-99 publication results.

## 21.4 Results

We find that our temporal Bayesian probabilities, with an adjusted initial prior, can mirror the Manifold Markets probabilities with two interesting divergences. While the initial probability starts at around 20% for both, the temporal Bayesian approach never goes above 30%. By contrast, the betting market, following the hype and virality of the LK-99 publication, reaches a peak at around 60%. Furthermore, while the betting market has a long tail of low probability starting in mid-August 2023, our approach more quickly disregards the hypothesis based on a continuing accumulation of studies showing that the LK-99 findings could not be replicated. Our temporal Bayesian model with the adjusted initial prior reaches a probability below 1% by mid-September 2023, but never entirely dismisses the chance that the hypothesis is true. Figure 29 showcases these results.

While we specifically select initial probabilities of 20% and 50%, both trajectories end with a steadily shrinking probability of accepting the hypothesis. Our initial prior probability of 50% mimics an unbiased observer with no knowledge about whether the hypothesis should be accepted or rejected. A prior probability of 20% could be considered a reasonable guess for an observer biased by a baseline understanding and suspicion of the claims and their corresponding evidence. Such an initial guess is admittedly subjective, as is the degree to which new information affects one’s inherent biases. We treat these settings as presets, however these are trivial for others to configure and assess based on their background. Finally, for claims with established scientific consensus, our approach is guaranteed to asymptotically approach that consensus where the rate of convergence varies according to these initial presets.

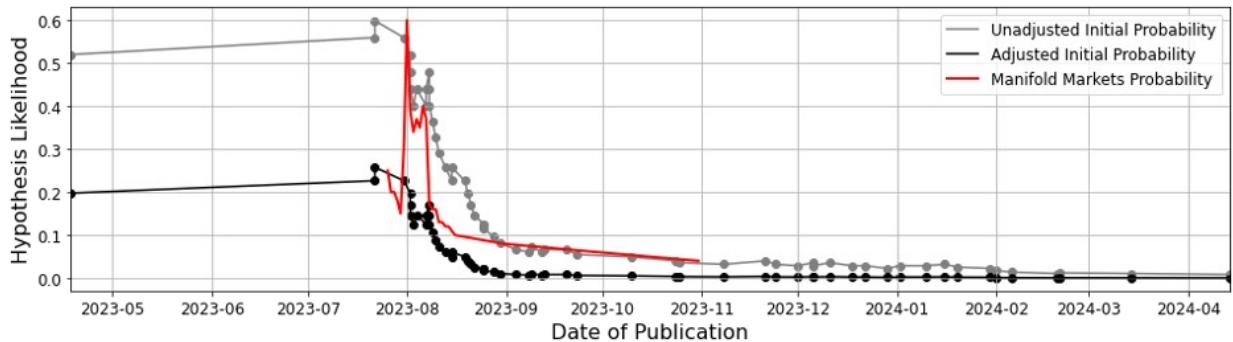


Figure 29: Likelihood of accepting the hypothesis “LK-99 is a room-temperature superconductor” via three approaches, from April 15, 2023 to April 15, 2024. The unadjusted initial probability (set to 50%) is shown in gray, the adjusted initial probability (set to 20%) is shown in black, and the probability according to the Manifold Markets online betting platform is shown in red.

## 21.5 Conclusion

Carefully validating a claim based a body of scientific literature can be a time-consuming and challenging prospect, especially without domain expertise. In this work, we demonstrate how a claim might be evaluated using a temporal Bayesian model based on a literature evaluation using natural language inference.

We show how our aggregated literature predictions allow us to quickly reject the hypothesis that LK-99 is a room-temperature superconductor. In the future, we hope to apply this approach to other scientific claims, including those with debates that are ongoing and as well as claims have an established scientific consensus. In general, we hope this approach will allow a researcher to quickly measure the scientific community's confidence in a claim, as well as aid the public in assessing both the veracity of a claim and the change in confidence driven by continued experimentation and observation.

## References

- [1] Settles, Burr. "Active learning literature survey." (2009).
- [2] Lehmann, Erich Leo, Joseph P. Romano, and George Casella. *Testing statistical hypotheses*. Vol. 3. New York: Springer, 1986.
- [3] Popper, Karl. *The logic of scientific discovery*. Routledge, 2005.
- [4] Lee, Sukbae, Ji-Hoon Kim, and Young-Wan Kwon. "The first room-temperature ambient-pressure superconductor." *arXiv preprint arXiv:2307.12008* (2023).
- [5] Chang, Kenneth. "LK-99 Is the Superconductor of the Summer." *New York Times* (2023).
- [6] Garisto, Dan. "Claimed superconductor LK-99 is an online sensation—But replication efforts fall short." *Nature* 620, no. 7973 (2023): 253-253.
- [7] Liu, Hanmeng, Leyang Cui, Jian Liu, and Yue Zhang. "Natural language inference in context-investigating contextual reasoning over long texts." In *Proceedings of the AAAI conference on artificial intelligence*, vol. 35, no. 15, pp. 13388-13396. 2021.
- [8] Murphy, Kevin P. "Conjugate Bayesian analysis of the Gaussian distribution." def 1, no. 2 $\sigma$  2 (2007): 16.
- [9] "Will the LK-99 room temp superconductivity pre-print replicate in 2023". Manifold Markets (2024). <https://manifold.markets/Ernie/will-the-lk99-room-temp-ambient-pre-17fc7cb7a2a0>

## 22 Multi-Agent Hypothesis Generation and Verification through Tree of Thoughts and Retrieval Augmented Generation

Authors: Aleyna Beste Ozhan, Soroush Mahjoubi

### 22.1 Introduction

Our project, developed during the “LLM Hackathon for Applications in Materials and Chemistry,” aims to accelerate scientific hypotheses and enhance the creativity of scientific inquiry. We propose using a multi-agent system of specialized large language models (LLMs) to streamline and enrich hypothesis generation and verification in materials science. This approach leverages diverse, fine-tuned LLM agents collaborating to generate and validate novel hypotheses more effectively. While similar pipelines have been proven useful in the social sciences [1], to the best of our knowledge, this work marks the first adaptation of such an approach to hypothesis generation in materials science. As illustrated in Figure 30, the system includes agents such as a background provider, an inspiration generator, a hypothesis generator, and three evaluators. Each agent plays a crucial role in formulating and assessing hypotheses, ensuring only the most viable and compelling ideas are developed. This innovative approach fosters an environment conducive to scientific inquiries.

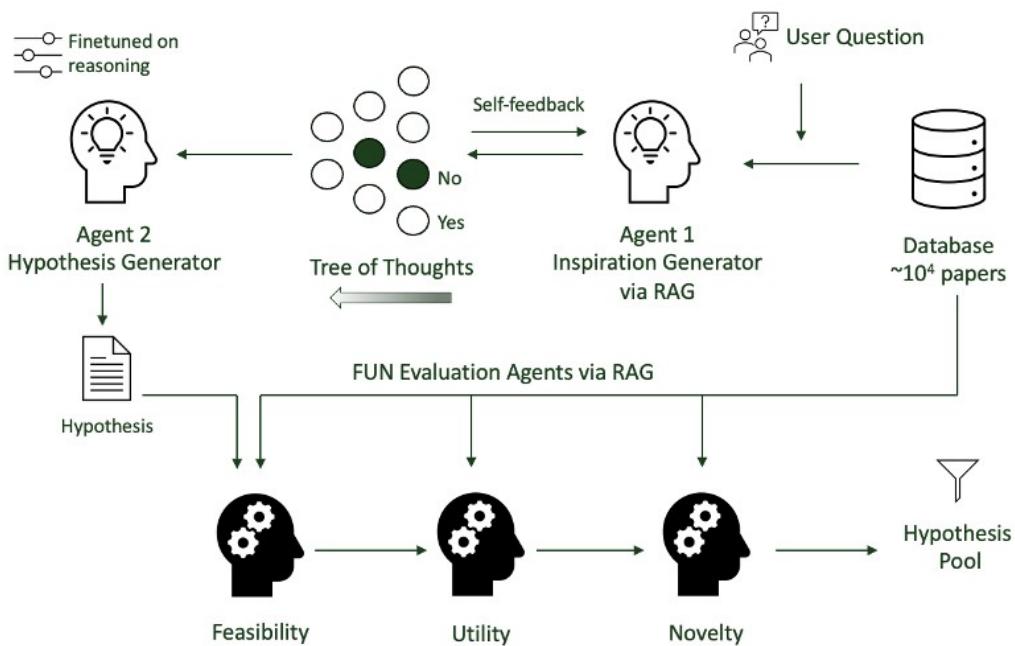


Figure 30: Multi-Agent Hypothesis Generation and Verification Pipeline

### 22.2 Methodology

**Background Extraction** The background extraction module is designed to search through a vast database for relevant information directly related to the user’s query. This module employs advanced embedding-based retrieval techniques to identify and extract pertinent corpus. As new papers and findings are added to the repository, the system dynamically updates, ensuring the use of the most current and relevant information.

**Inspiration Generator Agent** The inspiration generator agent leverages extensive background data to effectively formulate inspirations using a Retrieval Augmented Generation (RAG) mechanism. Serving as

the strategic core of the hypothesis generation process, it draws inspiration from a broad spectrum of sources to spawn diverse hypotheses, similar to the branching structure of a "Tree of Thoughts (ToT)" [2]. The agent samples "k" candidates as possible solutions, evaluates their effectiveness through self-feedback and votes on the most promising candidates. The selection is narrowed down to "b" promising options per step, with this structured approach helping the agent systematically refine its solutions.

**Hypothesis Generator** Based on the background information and the inspirations, this module generates meaningful research hypotheses. It is fine-tuned on reasoning datasets, such as Atlas, which encompasses various types of reasoning including deductive and inductive reasoning, cognitive biases, decision theory, and argumentative strategies [3].

### 22.3 Evaluator Agents

Once a hypothesis is generated, a RAG mechanism is used to fetch relevant abstracts from the dataset to evaluate the hypothesis. The evaluation is based on three critical aspects adopted from existing literature for the materials science domain:

*Feasibility:* Assesses whether the hypothesis is realistically grounded and achievable based on current scientific knowledge and technology.

*Utility:* Evaluates the practical value of the hypothesis, considering its potential to solve problems, enhance experimental design, or lead to beneficial exploratory paths.

*Novelty:* Measures the uniqueness and originality of the hypothesis, encouraging the generation of innovative ideas that advance scientific understanding.

### 22.4 Case Study

Our case study focuses on sustainable practices in concrete design at the material supply level. We processed 66,000 abstracts related to cement and concrete, converting them into a Chroma vector database using the sentence transformer all-MiniLM-L6-v2 [4]. This model maps abstracts to a 384-dimensional dense vector space. We then queried the embedding-based retrieval system with questions related to material-level solutions for sustainability in concrete, retrieving 10,000 relevant abstracts.

*Example Query:* "How can we incorporate industrial wastes and byproducts as supplementary cementitious materials to promote sustainability while maintaining its fresh and hardened properties such as strength?"

### 22.5 Results

In the initial phase of our "Tree of Thoughts" structure, we generated approximately 5,000 inspirations. These inspirations were refined to around 1,000 through a distillation step. The hypothesis generator, GPT-3.5 Turbo, fine-tuned on 13,000 data points from the AtlasUnified/Atlas-Reasoning dataset, produced one hypothesis per inspiration. The evaluation process involved three agents assessing feasibility, utility, and novelty (FUN) using embedding-based retrieval to identify relevant abstracts. For enhanced precision, GPT-4 was employed during the evaluation stages. Ultimately, hypotheses that withstood all evaluation stages were included in the hypothesis pool. Out of the initial 1,000 hypotheses, 243 passed the feasibility filter, 175 were deemed useful, and only 12 were found to be highly novel. The 12 hypotheses deemed feasible, novel, and useful are listed in [Table 5](#).

### 22.6 Future Directions: Adaptability to Other Material Systems and Cross-Domain Applications

To apply it to another material system, the first background query would be modified to target the new material of interest, and a database relevant to that material system would be employed. Also, this framework is not limited to materials science; it can be applied across various domains. For example, ideas generated from civil engineering could inspire hypotheses in materials science. A background provider querying civil engineering databases might produce inspirations that, when evaluated by our multi-agent system, lead to

innovative hypotheses in materials science. Similarly, within the domain of materials science, inspirations generated based on concrete research could be used to develop hypotheses for other materials, such as ceramics or composites. This cross-pollination of ideas can foster creativity and drive breakthroughs by applying concepts from one domain to another.

Table 5: Final hypothesis pool for the study of Section 22

No.	Hypothesis
<b>1</b>	Incorporating Stainless Steel (SS) micropowder from additive manufacturing into cement paste mixtures can improve the mechanical strength and durability of the mixture, with an optimal addition of 5% SS micropowder by volume.
<b>2</b>	The use of synthesized zeolites in self-healing concrete can significantly improve the durability and longevity of concrete structures.
<b>3</b>	The utilization of municipal solid waste (MSW) in cement production by integrating anaerobic digestion and mechanical-biological treatment to produce refuse-derived fuel (RDF) for cement kilns can reduce environmental impacts, establish a sustainable waste-to-energy solution, and create a closed-loop process that aligns waste management with cement production for a more sustainable future.
<b>4</b>	The use of smart fiber-reinforced concrete systems with embedded sensing capabilities can revolutionize infrastructure monitoring and maintenance by providing real-time feedback on structural health, leading to safer and more resilient built environments.
<b>5</b>	The use of advanced additives or nanomaterials in geopolymers cement can enhance its mechanical properties and durability, leading to more reliable CO <sub>2</sub> sequestration projects.
<b>6</b>	The use of carbonated steel slag as an aggregate in concrete can enhance the self-healing performance of concrete, leading to improved durability and longevity.
<b>7</b>	The synergistic effect of combining different pozzolanic materials with varying particle sizes and reactivities can lead to the development of novel high-performance concrete formulations with superior properties.
<b>8</b>	Smart bio concrete incorporating bacterial silica leaching exhibits superior strength, durability, and reduced water absorption capacity compared to traditional concrete.
<b>9</b>	Novel eco-concrete formulation developed by combining carbonated-aggregates with other sustainable materials like volcanic ash or limestone powder can create a carbon-negative concrete with superior mechanical strength, durability, and thermal conductivity.
<b>10</b>	The use of nano-enhanced steel fiber reinforced concrete (NSFRC) will result in a significant improvement in the mechanical properties, durability, and crack resistance of concrete structures compared to traditional steel fiber reinforced concrete.
<b>11</b>	The combined addition of silica fume (SF) and nano-silica (NS) can further enhance the sulphate and chloride resistance to higher than possible with the single addition of SF or NS.
<b>12</b>	The utilization of oil shale fly ash (OSFA) in concrete production can be optimized to develop sustainable and high-performance construction materials.

## References

[1] Yang, Zonglin, et al. "Large Language Models for Automated Open-domain Scientific Hypotheses Discovery." arXiv preprint [arXiv:2309.02726](https://arxiv.org/pdf/2309.02726.pdf) (2023). [https://arxiv.org/pdf/2309.02726](https://arxiv.org/pdf/2309.02726.pdf)

- [2] Yao, Shunyu, et al. "Tree of thoughts: Deliberate problem solving with large language models." Advances in Neural Information Processing Systems 36 (2024).
- [3] <https://huggingface.co/datasets/AtlasUnified/Atlas-Reasoning/commits/main>
- [4] <https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

## 23 ActiveScience

Authors: *Min-Hsueh Chiu*

### 23.1 Introduction

Humans have conducted material research for thousands of years, yet the vast chemical space, estimated to encompass up to  $10^{60}$  compounds, remains unexplored. Traditional research methods often focus on incremental improvements, making the discovery of new materials a slow process. As data infrastructure has developed, data mining techniques have increasingly accelerated material discovery. However, three significant obstacles hinder this process. First, the availability and sparsity of data present a major challenge. Comprehensive and high-quality datasets are essential for effective data mining, yet materials science suffer from limited data availability. Second, each database typically consists of specific types of quantitative properties, which may not fully meet researchers' needs. This fragmentation and specialization of databases can impede the holistic analysis necessary for breakthrough discoveries. Third, scientists usually focus on certain materials and related articles, decreasing the likelihood of deeply exploring diverse literature that reports potential materials or applications not yet utilized in their specific field. This siloed approach further limits the scope of discovery and innovation.

Unlike the intrinsic properties found in databases, scientific articles provide unstructured but higher-level information, such as applications, material categories, and properties that might not be explicitly recorded in databases. Additionally, these texts include inferences and theories proposed by domain experts, which are crucial for guiding research directions. The challenge lies in automatically extracting and digesting this unstructured text into actionable insights. This process typically requires experienced experts, creativity, and a measure of luck to identify the next desirable candidate. These challenges motivate the potential of utilizing large language models to parse scientific reports and integrate the extracted information into knowledge graphs, thereby constructing high-level insights.

### 23.2 Approach

The Python-based ActiveScience framework consists of three key functionalities: data source API, large language model, and graph database. LangChain is employed for downstream applications within this framework. The schematic pipeline is illustrated in Figure 1. Notably, this framework is not restricted to the specific packages or APIs used in this demonstration; alternative tools that provide the required functionality and input/output can also be integrated.

ArXiv APIs were used to retrieve scientific report titles, abstracts, and URLs. This demonstration focused on reports related to alloys. Consequently, the string "cat:cond-mat.mes-hall AND ti:alloy" was queried in the ArXiv APIs, which returned the relevant articles. GPT-3.5 Turbo models was used to access the large language model. The system's role was defined as "You are a material science professor and want to extract information from the paper's abstract." The provided prompt was: "Given the abstract of a paper, can you generate a Cypher code to construct a knowledge graph in Neo4j? ..." along with the designated ontology schema. The generated Cypher code is then input into Neo4j, which is used to ingest the entity relationships, store the knowledge graph, and provide robust visualization and querying interfaces.

### 23.3 Results

With the implemented knowledge graph, GraphCypherQACChain module from LangChain was employed to perform retrieval-augmented generation. For instance, when asked, "Give me the top 3 references URLs where the Property contains 'opti'?" GraphCypherQACChain automatically generates a Cypher query according to the designated schema, executes it in Neo4j, and ultimately returns the relevant answer, as shown in the right bottom box in Figure 31. Although this demonstration used a simple question, more complex queries can be processed using this framework. However, handling such queries effectively will require more refined prompting techniques.

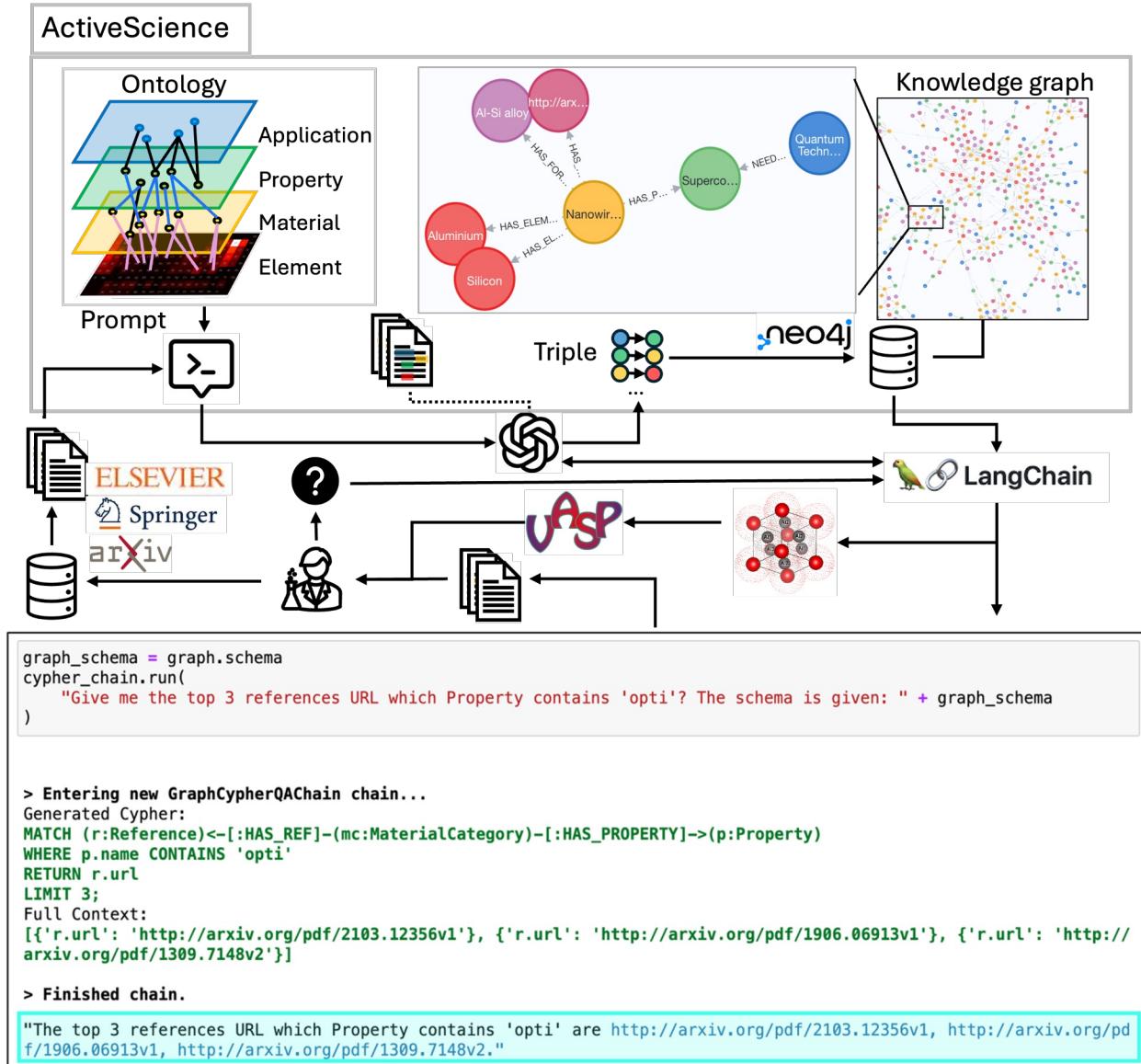


Figure 31: A Schematic illustration of ActiveScience architecture and its potential applications. Code snippet demonstrating the use of LangChain.

## 23.4 Conclusion and future works

The work demonstrates the pipeline of integrating large language model, knowledge graph, and a question-answering interface. Several aspects of this framework can be enhanced. Firstly, using domain-specific language models in the materials domain could improve entity and relationship recognition. Secondly, enhancing entity resolution and data normalization could lead to a more concise and informative knowledge graph, thereby improving the quality of answers. Thirdly, designing more effective prompting strategies, such as chain-of-thought prompting, could enhance the quality of both answers and code generation.

## 23.5 Data and Code Availability

All code and data used in this study are publicly available in the following GitHub repository: <https://github.com/minhsueh/ActiveScience>

## 24 G-Peer-T: LLM Probabilities For Assessing Scientific Novelty and Nonsense

Authors: Alexander Al-Feghali, Sylvester Zhang

Large language models (LLMs) and foundation models have garnered significant attention lately due to their natural language programmability and potential to parse high-dimensional data from reactions to the scientific literature [1]. While these models have demonstrated utility in various chemical and materials science applications, we propose leveraging their designed strength in language processing to develop a first pass peer review system for materials science research papers [2].

Traditional-gram tests such as BLEU or ROUGE, as well as X-of-thought LLM-based evaluations, are not sensitive enough for creativity or diversity in scientific writing [3]. Our approach utilizes the learned probabilistic features to establish a baseline for typical scientific language in materials science, based on fine-tuning on materials science abstracts through a historical train-test split. New abstracts are scored by their weighted-average probabilities, identifying those that deviate from the expected norms, flagging both possibly innovative or potentially nonsensical works.

As a proof-of-concept in this direction, we fine-tuned two models: OPT (6.7B) and TinyLLama (1.1B) using the Huggingface PEFT library's Low Rank Adapters (LoRA) to access the log probabilities of the abstracts, which is not typically accessible for modern API services [4–6]. Our results come with the usual caveats for small models with small computational costs.

We curated a dataset of 6000 abstracts from PubMed, published between 2017–2020, focusing on Materials Science and Chemistry [7]. The models were fine-tuned over 200 steps using this dataset. We compared highly cited papers (>200 citations) with those of average citation counts. Our preliminary findings suggest that higher-cited papers exhibit less “typical” language use, with mean log probabilities of  $-2.24 \pm 0.32$  for highly cited works compared to  $-1.79 \pm 0.3$  for average papers. However, the calculated p-value of 0.07 indicates that these results are not statistically significant at the conventional 0.05 level.

Full training with more steps on larger models, as well as more experimentation and method optimization, would yield more reliable results and be of modern relevance. Our documented code with step-by-step instructions is available in the repository [8].

## References

- [1] K. M. Jablonka, P. Schwaller, A. Ortega-Guerrero, B. Smit, *Nat. Mach. Intell.*, 2024, 6, 161–169.
- [2] D. A. Boiko, R. MacKnight, B. Kline, G. Gomes, *Nature*, 2023, 624, 570–578.
- [3] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, C. Zhu, *Proc. 2023 Conf. Empir. Methods Nat. Lang. Process.*, 2023, 2511–2522.
- [4] S. Mangrulkar, S. Gugger, L. Debut, Y. Belkada, S. Paul, B. Bossan, *PEFT: State-of-the-art Parameter-Efficient Fine-Tuning Methods*; GitHub: <https://github.com/huggingface/peft>, 2022.
- [5] P. Zhang, G. Zeng, T. Wang, W. Lu, *TinyLlama: An Open-Source Small Language Model*; arXiv:2401.02385.
- [6] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. V. Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. S. Koura, A. Sridhar, T. Wang, L. Zettlemoyer, *OPT: Open Pre-trained Transformer Language Models*; arXiv:2205.01068.
- [7] National Center for Biotechnology Information (NCBI) [Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; <https://www.ncbi.nlm.nih.gov/> 1998.
- [8] A. Al-Feghali, S. Zhang, *G-Peer-T* ; GitHub: <https://github.com/alfegh/G-Pee> 2024.

## 25 ChemQA: Evaluating Chemistry Reasoning Capabilities of Multi-Modal Foundation Models

Authors: Ghazal Khalighinejad, Shang Zhu, Xuefeng Liu

### 25.1 Introduction

Current foundation models exhibit impressive capabilities when prompted with text and image inputs in the chemistry domain. However, it is essential to evaluate their performance on text alone, image alone, and a combination of both to fully understand their strengths and limitations. In chemistry, visual representations often enhance comprehension. For instance, determining the number of carbons in a molecule is easier for humans when provided with an image rather than SMILES annotations. This visual advantage underscores the need for models to effectively interpret both types of data. To address this, we propose ChemQA—a benchmark dataset containing problems across five question-and-answering (QA) tasks. Each example is presented with isomorphic representations: visual (images of molecules) and textual (SMILES). ChemQA enables a detailed analysis of how different representations impact model performance.

We observe from our results that models perform better when given both text and visual inputs compared to when they are prompted with image-only inputs. Their accuracy significantly decreases when provided with only visual information, highlighting the importance of multimodal inputs for complex reasoning tasks in chemistry.

### 25.2 ChemQA: A Benchmark Dataset for Multi-Modal Chemical Understanding

Inspired by the existing work of IsoBench [2] and ChemLLMBench [3], we create a multimodal question-and-answering dataset on chemistry reasoning, ChemQA [5] containing five QA tasks:

- **Counting Numbers of Carbons and Hydrogens in Organic Molecules:** adapted from the 600 PubChem molecules created by [3], evenly divided into validation and evaluation datasets.
  - Example: Given a molecule image or its SMILES notation, identify the number of carbons and hydrogens.
- **Calculating Molecular Weights in Organic Molecules:** adapted from the 600 PubChem molecules created by [3], evenly divided into validation and evaluation datasets.
  - Example: Given a molecule image or its SMILES notation, calculate its molecular weight.
- **Name Conversion: From SMILES to IUPAC:** adapted from the 600 PubChem molecules created by [3], evenly divided into validation and evaluation datasets.
  - Example: Convert a given SMILES string or a molecule image to its IUPAC name.
- **Molecule Captioning and Editing:** inspired by [3], adapted from the dataset provided in [1], following the same training, validation, and evaluation splits.
  - Example: Given a molecule image or its SMILES notation, find the most relevant description of the molecule.
- **Retro-synthesis Planning:** inspired by [3], adapted from the dataset provided in [4], following the same training, validation, and evaluation splits.
  - Example: Given a molecule image or its SMILES notation, find the most likely reactants that can produce the molecule.

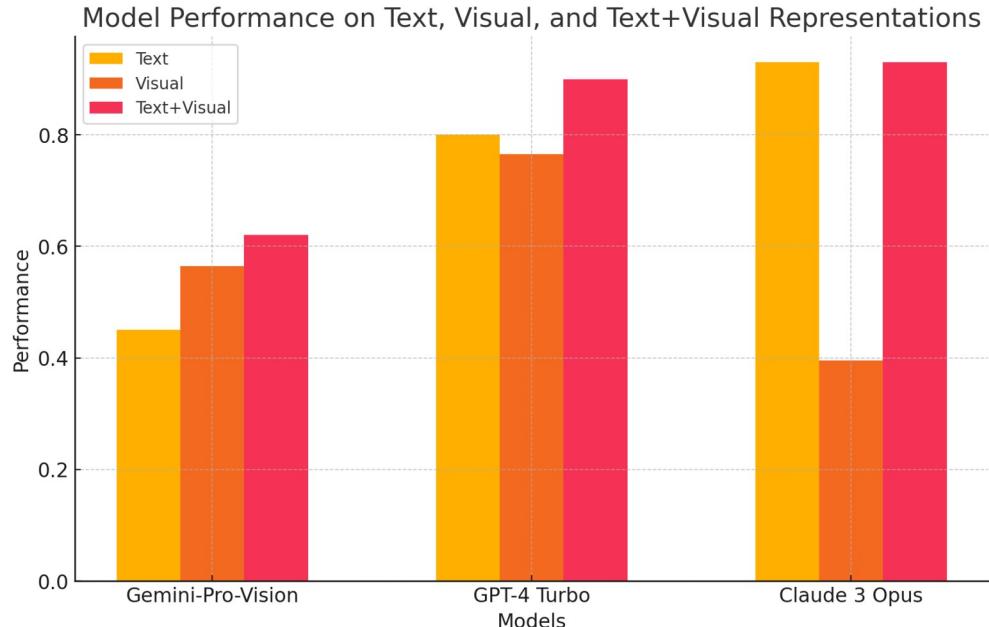


Figure 32: Performance of Gemini Pro, GPT-4 Turbo, and Claude3 Opus on text, visual, and text+visual representations. The plot shows that models achieve higher accuracy with combined text and visual inputs compared to visual-only inputs.

### 25.3 Evaluating Foundation Model Performances on ChemQA

The model evaluation results are shown in Figure 32. According to the plot, models perform better with text and visual inputs combined, while their accuracy drops when given only visual inputs. Claude performs well in text-only tasks, whereas Gemini and GPT-4 Turbo perform better with visual or combined inputs. This highlights the importance of evaluating models with different input modalities to understand their full capabilities and limitations.

### 25.4 Conclusion and Future Work

The evaluation of multimodal language models in the chemistry domain demonstrates that the integration of both text and visual inputs significantly enhances model performance. This suggests that for complex reasoning tasks in chemistry, multimodal approaches are more effective than relying on a single type of input. Future work should continue to explore and refine these multimodal strategies to further improve the accuracy and applicability of AI models in specialized fields such as chemistry.

## References

- [1] Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between molecules and natural language. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 375–413, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.26. URL <https://aclanthology.org/2022.emnlp-main.26>
- [2] Deqing Fu, Ghazal Khalighinejad, Ollie Liu, Bhuwan Dhingra, Dani Yogatama, Robin Jia, and Willie Neiswanger. Isobench: Benchmarking multimodal foundation models on isomorphic representations, 2024. URL <https://arxiv.org/abs/2404.01266>

- [3] Taicheng Guo, Kehan Guo, Bozhao Nan, Zhenwen Liang, Zhichun Guo, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. What can large language models do in chemistry? a comprehensive benchmark on eight tasks, 2023. URL <https://arxiv.org/abs/2305.18365>
- [4] Ross Irwin, Spyridon Dimitriadis, Jiazen He, and Esben Jannik Bjerrum. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1):015022, January 2022. doi: 10.1088/2632-2153/ac3ffb. URL <https://dx.doi.org/10.1088/2632-2153/ac3ffb>
- [5] Shang Zhu, Xuefeng Liu, and Ghazal Khalighinejad. ChemQA: a multimodal question-and-answering dataset on chemistry reasoning. <https://huggingface.co/datasets/shangzhu/ChemQA>

## 26 LithiumMind - Leveraging Language Models for Understanding Battery Performance

Authors: Xinyi Ni, Zizhang Chen, Rongda Kang, Sheng-Lun Liao, Pengyu Hong, Sandeep Madireddy

### 26.1 Introduction

In this project, we explore multiple applications of Large Language Models (LLMs) in analyzing the performance of lithium metal batteries. Central to our investigation is the concept of Coulombic efficiency (CE), a critical measure in battery technology that assesses the efficiency of electron transfer during a battery's charge and discharge cycles. Improving the CE is essential for advancing the adoption of high-energy-density lithium metal batteries, which are crucial for next-generation energy storage solutions. A key focus of our analysis is the role of the liquid electrolyte engineering, which is instrumental in determining the battery's cyclability and improving the CE.

We explored two methods of integrating LLM as a supplemental tool in battery research. First, we utilize LLMs as information extractors to distill structural knowledge essential for the design of electrolytes from the vast amount of paper. Second, we introduce an LLM-powered chatbot specifically tailored to respond to inquiries related to lithium metal batteries. We believe these applications of LLMs may potentially enhance our capabilities in battery innovation, streamlining research processes and increasing the accessibility of specialized knowledge. Our code is available at: <https://github.com/KKbeckang/LiGPT-Beta>

### 26.2 Investigation 1 Structural Knowledge Extraction

**Motivation** In this project, we utilize LLM to extract CE-electrolyte pairs from over 70 papers. A recent study [1] (Dataset S01) provides a reliable database about the relationship between Coulombic Efficiency (CE) and battery electrolytes, along with a list of relevant papers. The data was filtered and cleaned by human experts. We managed to automate the data collection procedure through LithiumMind and aimed to discover more relationships from the paper list. The pipeline is described in the left column of Figure 1.

**Method** The pipeline consists of the following steps:

- **Parse PDF:** The raw papers were saved as PDF files and loaded into text using a PDF parser.
- **Retrieve Relevant Context:** Instead of extracting information directly from the whole paper, we ingest the documents into a vector datastore. We combined three domain-specific embedding models - MaterialBERT, ChemBERT, and OpenAI - together as a powerful retriever through the LangChain LOR module. The retriever finds 10 text chunks for each paper that are most relevant to Coulombic Efficiency.
- **CE Extraction:** We defined the schema for extracting key information, including Coulombic Efficiency and the solvent and salt of the electrolyte. We provided few-shot in-context instruction to GPT-4 Turbo JSON mode. The extracted output is saved in JSON format.

**Preliminary Results** The LLM extracted 334 CE-electrolyte pairs in 71 papers, while the original paper found 152 pairs. Since it is difficult to verify our results without the help of human experts, we filtered the extracted pairs through the Coulombic Efficiency and found 46 matches to the original dataset. We classified these verifiable data into three categories: *correct* (the extracted data exactly matches the human-labeled data); *incorrect* (the types or amounts of the solvent/salt do not match the labels); and *unknown* (the extracted data provides more details than the human-labeled data). For example, the label only shows the types of the salts, but the extracted data contains not only the correct types but also the mix ratio of different salts. The results are shown in Table 6:

Table 6: Results

Correct	Incorrect	Unknown
72.4%	23.4%	4.2%

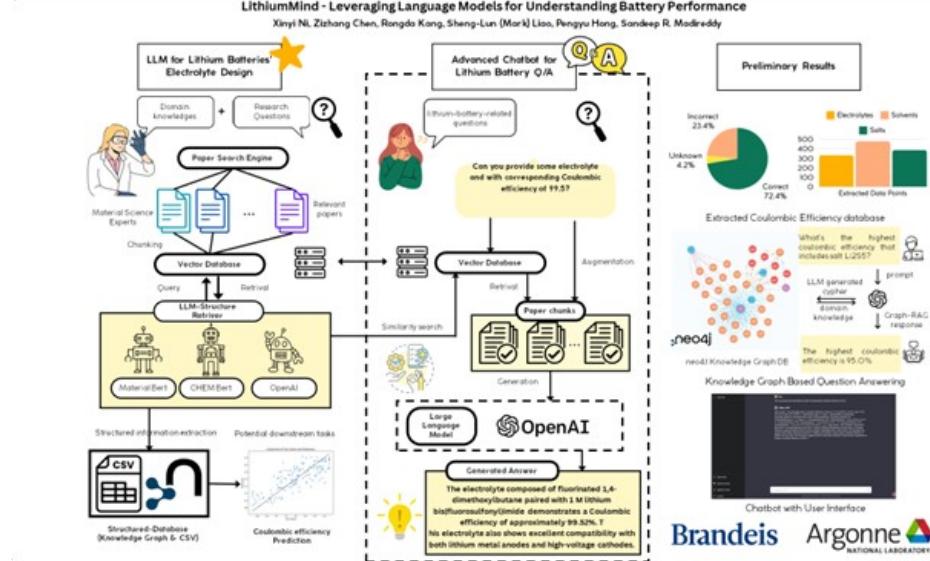


Figure 33: Summary of our LLM hackathon project.

**Future Work** One major challenge is the low recall of the extracted information, as only 46 out of 152 labeled pieces of information were retrieved. Upon investigating the papers, we found that much of the Coulombic Efficiency was recorded in tables and figures, which were dropped during PDF parsing. It is necessary to introduce multimodal LLMs to further investigate those papers.

### 26.3 Investigation 2Advanced Chatbot for Lithium Battery Q/A

In this exploration, we utilized a curated collection of research papers focused on lithium metal batteries to construct a vector database and developed a chatbot employing a Retrieval-Augmented Generation (RAG) framework. Our Q/A pipeline includes two types of answer strategies: General Q/A and Knowledge-graph-based Q/A. The pipeline is described in Figure 1.

**General Q/A Building** In this section, we detail our comprehensive Q/A pipeline designed for the exploration of lithium metal battery-related research. We began by sourcing and downloading 71 research papers pertinent to lithium metal batteries. The information extracted from these papers was encoded into vectors and stored in Chroma databases. To best reflect the specialized language of the field, we created two distinct databases: one utilizing MaterialsBERT for materials science content, and another using ChemBERT for chemical context.

During the retrieval phase, we employ LOTR (MergerRetriever) to enhance the precision of document retrieval. Upon receiving a user query, the system retrieves relevant document segments from each database. It then removes any redundant results from the merged outputs and selects the top 10 most pertinent document chunks. Finally, both the selected context and the user query are processed by GPT-4 Turbo to generate an informed and accurate response. This pipeline exemplifies a robust integration of state-of-the-art technologies to facilitate advanced research interrogation and knowledge discovery in the domain of lithium metal batteries.

**Knowledge Graph-based Q/A Building** We built a knowledge graph with the extracted information using Neo4j. The knowledge graph consists of four node types:

- **Solvent node:** properties include name, SMILES, density, and weight.
- **Salt node:** properties include name, SMILES, and weight.
- **Electrolyte node:** properties include name and Coulombic Efficiency.
- **Reference node:** properties include title, content, and page number and, three edge types:
  - **(electrolyte)-[:VOLUME]->(solvent)**
  - **(electrolyte)-[:CONCENTRATION]->(salt)**
  - **(electrolyte)-[:CITED]->(reference)**

The built knowledge graph can be accessed and visualized using the Neo4j web application.

**Knowledge Graph Enhanced Question Answering** By using GraphCypherQAChain in LangChain, LLMs can generate Cypher queries to solve users' questions. This capability allows the LLMs to address user queries by filtering and leveraging relationships between data points, which is particularly valuable in complex domains such as lithium battery technology. This integration ensures that our RAG pipeline is adept at handling domain-specific questions and excels in scenarios where understanding the interconnections within data is crucial.

## References

[1] Kim, S.C., et al. "Data-driven electrolyte design for lithium metal anodes." Proceedings of the National Academy of Sciences 120.10 (2023):e2214357120 <https://www.pnas.org/doi/full/10.1073/pnas.2214357120>

## 27 KnowMat: Transforming Unstructured MateriScience Literature into Structured Knowledge

Authors: Hasan M. Sayeed, Ramsey Issa, Trupti Mohanty, Taylor Sparks

### 27.1 Introduction

The rapid growth of materials science has led to an explosion of scientific literature, often presented in unstructured formats that pose significant challenges for systematic data extraction and analysis. To address this, we developed KnowMat, a novel tool designed to transform complex, unstructured material science literature into structured knowledge. Leveraging advanced Large Language Models (LLMs) such as GPT-3.5, GPT-4, Llama 3, KnowMat automates the extraction of critical information from scientific texts, converting them into structured JSON formats. This tool not only enhances the efficiency of data processing but also facilitates deeper insights and innovation in material science research. KnowMat's user-friendly interface allows researchers to input material science papers, which are then parsed and analyzed to extract key insights. The tool's versatility in handling multiple input files and its capacity for customization through sub-field specific prompts make it an indispensable resource for researchers aiming to streamline their workflow. Additionally, KnowMat's ability to integrate with other tools and platforms, along with its support for various LLMs, ensures that it remains adaptable to the evolving needs of the research community.

### 27.2 Method

**Data Parsing and Extraction** The KnowMat workflow begins with parsing unstructured text from material science literature using a tool called Unstructured [1]. This tool reads the input file, separates out the sections, and stores everything in a machine-readable format. This initial step involves identifying relevant sections and extracting pertinent information related to material compositions, properties, and experimental conditions.

**Customizable Prompts** KnowMat provides field-specific prompts for several fields, and it offers the flexibility for users to customize these prompts further or create their own prompts for new fields. This feature ensures that the extracted data is both relevant and comprehensive, tailored to the specific needs of the researcher. The interface allows users to define the scope and focus of the extraction process effectively.

**Integration and Interoperability** To enhance usability and interoperability, KnowMat supports seamless integration with other tools and platforms. Extracted results can be easily exported in CSV format, enabling straightforward data sharing and further analysis. The tool's flexibility extends to its compatibility with various LLMs, including both subscription-based models like GPT-3.5 and GPT-4 [2], and open-source models like Llama 3 [3]. This ensures that researchers can select the most suitable LLM for their specific requirements.

### 27.3 Preliminary Results

In initial tests, KnowMat demonstrated promising results in the efficiency and accuracy of data extraction from material science literature. The tool successfully parsed and structured information from multiple papers, converting complex textual data into actionable insights. For instance, in a study involving superconductors, KnowMat accurately extracted detailed information on compositions, critical temperatures, and experimental conditions, presenting them in a structured JSON format.

### 27.4 Future Development

Looking ahead, future developments for KnowMat include enhancing the editing capabilities for field-specific prompts, improving the handling of multiple input files in a single operation, and expanding output options to further enhance flexibility and usability. Continuous improvement and expansion of field-specific prompts will ensure that KnowMat remains a valuable tool for researchers across various domains of material science.

## KnowMat: Transforming Complexity into Clarity

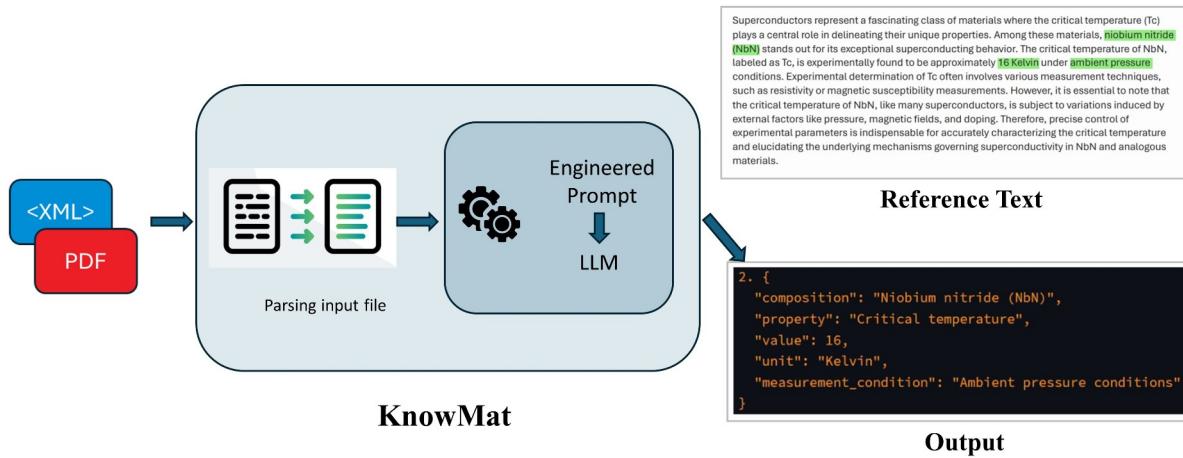


Figure 34: KnowMat Workflow. The graphical abstract illustrates the KnowMat workflow, which begins with parsing input files in XML or PDF formats. The Large Language Model (LLM) powered by engineered prompts processes the reference text to extract key information, such as material composition, properties, and measurement conditions, and converts it into a structured JSON output.

In conclusion, KnowMat represents a significant advancement in the field of knowledge extraction from scientific literature. By converting unstructured material science texts into structured formats, it provides researchers with powerful tools to unlock insights and drive innovation in their fields.

### 27.5 Data and code availability

[https://github.com/sparks-sayeed/LLMs\\_for\\_Materials\\_and\\_Chemistry\\_Hackathon](https://github.com/sparks-sayeed/LLMs_for_Materials_and_Chemistry_Hackathon)

### References

- [1] <https://unstructured.io/>
- [2] <https://platform.openai.com/docs/models>
- [3] <https://llama.meta.com/llama3/>

## 28 Ontosynthesis

Authors: Qianxiang Ai, Jiaru Bai, Kevin Shen, Jennifer D'Souza, Elliot Risch

### 28.1 Introduction

Organic synthesis is often described in unstructured text without a standard taxonomy, which makes synthesis data less searchable and less compatible with downstream data-driven tasks (e.g., retrosynthesis, condition recommendation) compared to structured records. The specificities and writing styles of these descriptions also vary, ranging from simple sentences about chemical transformations to long paragraphs that include procedure details. This leads to ambiguities, unidentified missing information, challenges in comparing different syntheses, and can impede proper experimental reproduction.

In last year's hackathon, we fine-tuned an open-source LLM to extract data structured in the Open Reaction Database schema from synthesis procedures [1]. While this method has proved to be successful for patent text, it relies on existing unstructured-structured data pairs and does not generalize well to non-patent writing styles. The dependency of fine-tuning on existing data makes it less useful, especially when considering new writing styles, or newly developed data structures or ontologies are preferred.

In this project, we explore the potential of LLMs in structured data extraction without fine-tuning. Specifically, given an ontology (formally defined concepts and relationships) for organic synthesis, we aim to extract structured information as Resource Description Framework (RDF) graphs from unstructured text using LLMs with zero-shot prompting. RDF is a World Wide Web Consortium (W3C) standard that serves as a foundation for the Semantic Web and expressing meaning and relationships between data. While LLMs can create ontologies on the fly for a given piece of text, "grounding" to a pre-specified ontology allows standardizing the extracted data and reasoning with existing axioms. The extraction workflow is wrapped in a web application which also allows visualization of the extracted results. We showcased the capability of our application with case studies where RDFs were extracted from reaction descriptions of different complexities and specificities.

### 28.2 Information extraction workflow

OpenAI's GPT model (gpt-4-turbo-2024-04-09) is used for structured information extraction through its ChatCompletion API. The prompt for an individual extraction task consists of two parts:

1. The System prompt: The given ontology in OWL format based on which the RDF graph is defined, along with supporting instructions on the task. The full prompt template is included in the project GitHub repository [2]. Two ontologies were used in this study:
  - (a) OntoReaction: A reaction ontology previously used in distributed self-driving laboratories [3];
  - (b) Synthesis Operation Ontology: A process chemistry ontology designed for common operations used in organic synthesis [4].
2. The User prompt: The unstructured text from which the RDF graph is extracted.

The final OpenAI API request is a combination of the User and System prompts:

```
from openai import OpenAI
client = OpenAI(api_key=api_key)
completion = client.chat.completions.create(
    model="gpt-4-turbo",
    messages=[{"role": "system", "content": system prompt, },
              {"role": "user", "content": unstructured data.text
    ],
)
```

Alternatively, one can use GPT assistants and provide the System prompt as a base instruction through OpenAI's web user interface.

## 28.3 Application

We collected a set of eight reaction descriptions taken from patents, journal articles (main text or supporting information), and lab notebooks. Each of them is assigned a complexity rating and a specificity rating using three-point scales. Based on these test cases we found our workflow was able to produce valid RDF graphs representing the chemical reactions, even for multi-step reactions including many elements. Expert inspections indicate the resulting RDF graphs better represent the unstructured text when OntoReaction is used as the target ontology compared to the larger Synthesis Operation Ontology (the latter contains more classes and object properties).

Since the extracted data is in RDF format, they can be readily visualized using interactive graph libraries. Using dash-cytoscape<sup>[5]</sup>, we created an interface application to the extraction workflow. The interface allows submitting unstructured text as input to the extraction workflow with a user-provided OpenAI API key, retrieving and interactively visualizing the extracted knowledge graph, as well as displaying the extracted RDF text. A file-based caching routine is used to store previous extraction results. All code and test cases are available in the project GitHub repository [2].

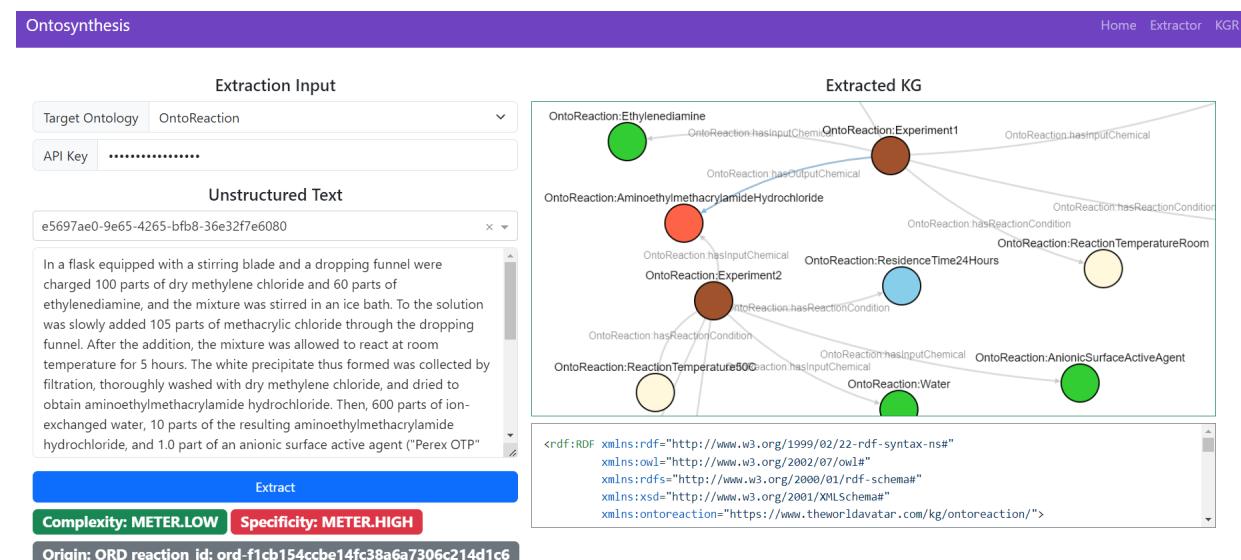


Figure 35

## 28.4 Acknowledgements

Q.A. acknowledges support by the National Institutes of Health under award number U18TR004149. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. J. D. acknowledges the SCINEXT project (BMBF, German Federal Ministry of Education and Research, Grant ID: 01IS22070).

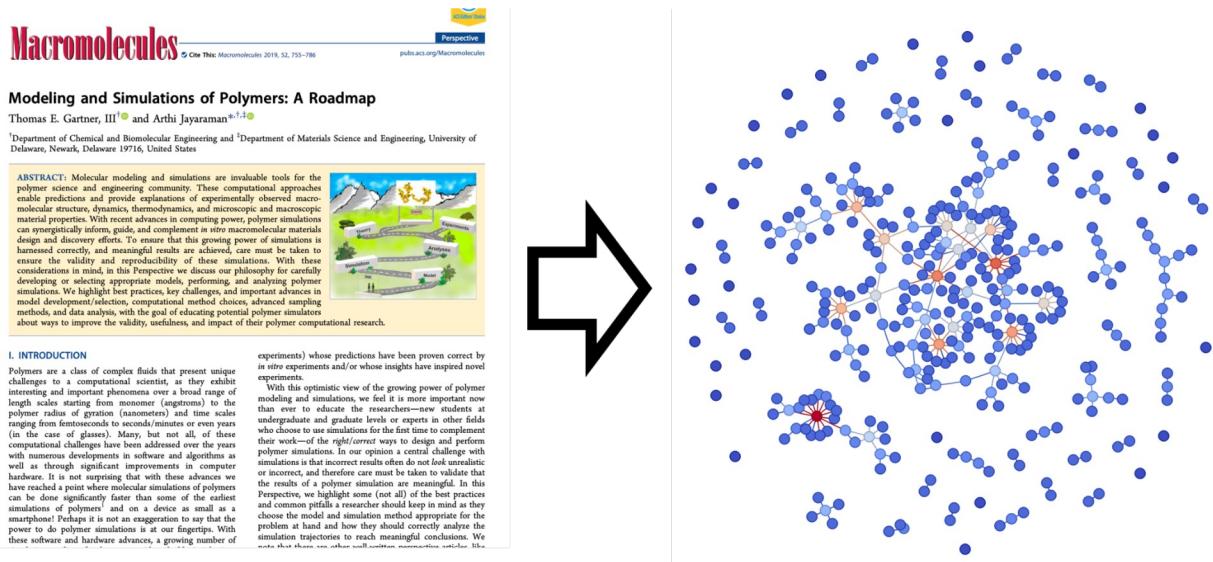
## References

- [1] Ai, Q.; Meng, F.; Shi, J.; Pelkie, B.; Coley, C. W. Extracting Structured Data from Organic Synthesis Procedures Using a Fine-Tuned Large Language Model. *ChemRxiv* April 8, 2024. <https://doi.org/10.26434/chemrxiv-2024-979fz>
- [2] Ontosynthesis, 2024. <https://github.com/qai222/ontosynth> (accessed 2024-07-07).
- [3] Bai, J.; Mosbach, S.; Taylor, C. J.; Karan, D.; Lee, K. F.; Rihm, S. D.; Akroyd, J.; Lapkin, A. A.; Kraft, M. A Dynamic Knowledge Graph Approach to Distributed Self-Driving Laboratories. *Nat. Commun.* 2024, 15 (1), 462. <https://doi.org/10.1038/s41467-023-44599-9>

- [4] Ai, Q.; Klein, C. Synthesis Operation Ontology. GitHub. <https://github.com/qai222/ontosynthesis/blob/main/ontologies/soo/soo.ttl> (accessed 2024-07-07).
- [5] Dash-Cytoscape: A Component Library for Dash Aimed at Facilitating Network Visualization in Python, Wrapped around Cytoscape.js. <https://dash.plotly.com/cytoscape> (accessed 2024-07-06).

## 29 Knowledge Graph RAG for Polymer Simulation

Authors: Jiale Shi, Weijie Zhang, Dandan Tang, Chi Zhang



### Document about Polymer Simulation

### Knowledge Graph RAG

Figure 36: Creating Knowledge Graph Retrieval-Augmented Generation (KGRAG) for Polymer Simulation.

Molecular modeling and simulations have become essential tools in polymer science and engineering, offering predictive insights into macromolecular structure, dynamics, and material properties. However, the complexity of polymer simulations poses challenges in model development/selection, computational method choices, advanced sampling techniques, and data analysis. While literature [1] provides guidelines to ensure the validity and reproducibility of these simulations, these resources are often presented in massive, unstructured text formats, making it difficult for new learners to systematically and comprehensively understand the correct approaches for model selection, simulation execution, and post-processing analysis. Therefore, this study presents a novel approach to address these challenges by implementing a Knowledge Graph Retrieval-Augmented Generation (KGRAG) system for building an AI chatbot focused on polymer simulation guidance and education.

Our team utilized the large language model GPT-3.5 Turbo [2] and Microsoft's GraphRAG [3, 4] framework to extract polymer simulation-related entities and relationships from unstructured documents, constructing a comprehensive KGRAG, as shown in Figure 36, where the nodes are colored by their degrees. Those nodes with high degrees include "Polymer Simulation," "Atomistic Model," "CG Model," "Force Field," and "Polymer Informatics," which are all the keywords about polymer simulation and modeling, illustrating the effective performance of entity extraction. We run the query engineering of KGRAG to ask questions. For comparative analysis, we also implemented a baseline Retrieval-Augmented Generation (RAG) system using LlamaIndex [5]. Upon comparing the responses from the baseline RAG and KGRAG by human experts, we found that the KGRAG demonstrates substantial improvements in question-answering performance when analyzing complex and high-level information about polymer simulation. This improvement is attributed to KGRAG's ability to capture semantic concepts and uncover hidden connections within the data, providing more accurate, logical, and insightful responses compared to traditional vector-based RAG methods.

This study contributes to the growing field of data-driven approaches in polymer science by offering a

powerful tool for knowledge extraction and synthesis. Our KGRAG system shows promise in enhancing the understanding of massive unstructured polymer simulation guidance in the relevant literature, potentially improving the validity and reproducibility of these polymer simulations, and accelerating the development of new polymer simulation methods. We found that the quality of prompts is crucial for effective entity extraction and knowledge graph construction. Therefore, our future work will focus on optimizing prompts for entity extraction, relationship building, and knowledge graph construction to further improve the system's performance and applicability in polymer simulation research.

## 29.1 Data availability

Repository: <https://github.com/shijiale0609/KG-RAG-LLM-Polymers>

## 29.2 Author

- Jiale Shi, Department of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States (jialeshi@mit.edu)
- Weijie Zhang, Department of Chemistry, University of Virginia, Charlottesville, Virginia 22904, United States (shv9px@virginia.edu)
- Dandan Tang, Department of Psychology, University of Virginia, Charlottesville, Virginia 22904, United States (gux8df@virginia.edu)
- Chi Zhang, Department of Automation Engineering, RWTH Aachen University, Aachen, North Rhine-Westphalia, 52062, Germany (chi.zhang2@rwth-aachen.de)

## References

- [1] Gartner, T. E., III; Jayaraman, A. Modeling and Simulations of Polymers: A Roadmap. *Macromolecules* 2019, 52 (3), 755-786. DOI: [10.1021/acs.macromol.8b01836](https://doi.org/10.1021/acs.macromol.8b01836)
- [2] GPT-3.5 Turbo. 2024. <https://platform.openai.com/docs/models/gpt-3-5-turbo> (accessed 2024/07/10).
- [3] Welcome to GraphRAG. 2024. <https://microsoft.github.io/graph> (accessed 2024/07/10).
- [4] Edge, D.; Trinh, H.; Cheng, N.; Bradley, J.; Chao, A.; Mody, A.; Truitt, S.; Larson, J. From Local to Global: A Graph RAG Approach to Query-Focused Summarization. 2024. (accessed 2024/07/10).
- [5] Llamaindex. 2024. <https://docs.llamaindex.ai/en/stable> (accessed 2024/07/10).

## 30 Synthetic Data Generation and Insightful Machine Learning for High Entropy Alloy Hydrides

Authors: Tapashree Pradhan, Devi Dutta Biswajeet

The generation of synthetic data in materials science and chemistry is traditionally performed by machine learning interatomic potentials (MLIPs) that approximate the first principle functional form used to compute wave functions of known electronic configurations [1]. Synthetic data is a component of the active-learning feedback loop that is utilized to retrain these potentials. The purpose of this active-learning component is to incorporate a wider range of physical conditions into the potential's application domain. However, the initial cost of data generation to train the MLIPs is a major setback for complex chemistries like in the case of high entropy alloys (HEAs).

The potential application of high entropy alloys in hydrogen storage [2] demands acceleration in the computation of surface and mechanical properties of the alloys by better approximation of the potential landscape. The use of Large Language Models (LLMs) in the generation of synthetic data to tackle this bottleneck problem poses an alternative to the traditional MLIPs. LLMs like ChatGPT, Llama3, Claude, and Gemini [3] learn from text-embeddings of the training data, capturing inherent trends between semantic or numerical relationships of the text and making them suitable for learning certain complex relationships in materials physics that might be present in the language itself.

The current work aims to build LLM applications working in conjunction with an external database of high entropy alloy hydrides via Retrieval-Augmented Generation (RAG) [4] to populate synthetic data for predictive modeling later. The inbuilt RAG feature of GPT-4 enables us to write a few prompts to make a synthetic data generator utilizing a custom dataset of HEA hydrides. The work also utilizes OpenAI's API [5] and the text-embedding-3-large model [6] to configure custom generators that can be fine-tuned via prompts for synthetic data generation.

The development of the entire product is aimed at a web-based application that allows users to upload their datasets and instruct the GPT model to generate more entries that can serve as training data for predictive ML models like hydrogen capacity predictors. The term "Insightful Machine Learning" refers to a sequential pipeline starting with (a) a reference database that serves as the retrieval component of an LLM, (b) the generation of synthetic data and features, and (c) getting insights from a chatbot on physics of the problem having multiple retrieval components inclusive of the predictive model. [Figure 37](#) shows the flowchart of the pipeline which is currently at the prototype stage under development. The current code to generate synthetic data is available for use and modification.

### 30.1 Future Directions

The proposed pipeline shown in Figure 1 requires a validation stage which is essential for active learning. The ongoing work involves the search for novel validation techniques taking inspiration from recently published works on information theory.

The future direction of the work is to complete the validation phase and have a product utilizing the pipeline for high entropy hydrides that can accelerate the search and discovery process without performing complex first principle calculations for data generation and training.

## References

- [1] Focassio, B., Freitas, M., Schleder, G.R. (2024). Performance assessment of universal machine learning interatomic potentials: Challenges and directions for materials' surfaces. *ACS Applied Materials & Interfaces*.
- [2] Marques, F., Balcerzak, M., Winkelmann, F., Zepon, G., Felderhoff, M. (2021). Review and outlook on high-entropy alloys for hydrogen storage. *Energy & Environmental Science*, 14(10), 5191-5227.
- [3] Jain, S.M. (2022). *Hugging face: Introduction to transformers for NLP with the hugging face library and models to solve problems*. Berkeley, CA: Apress.

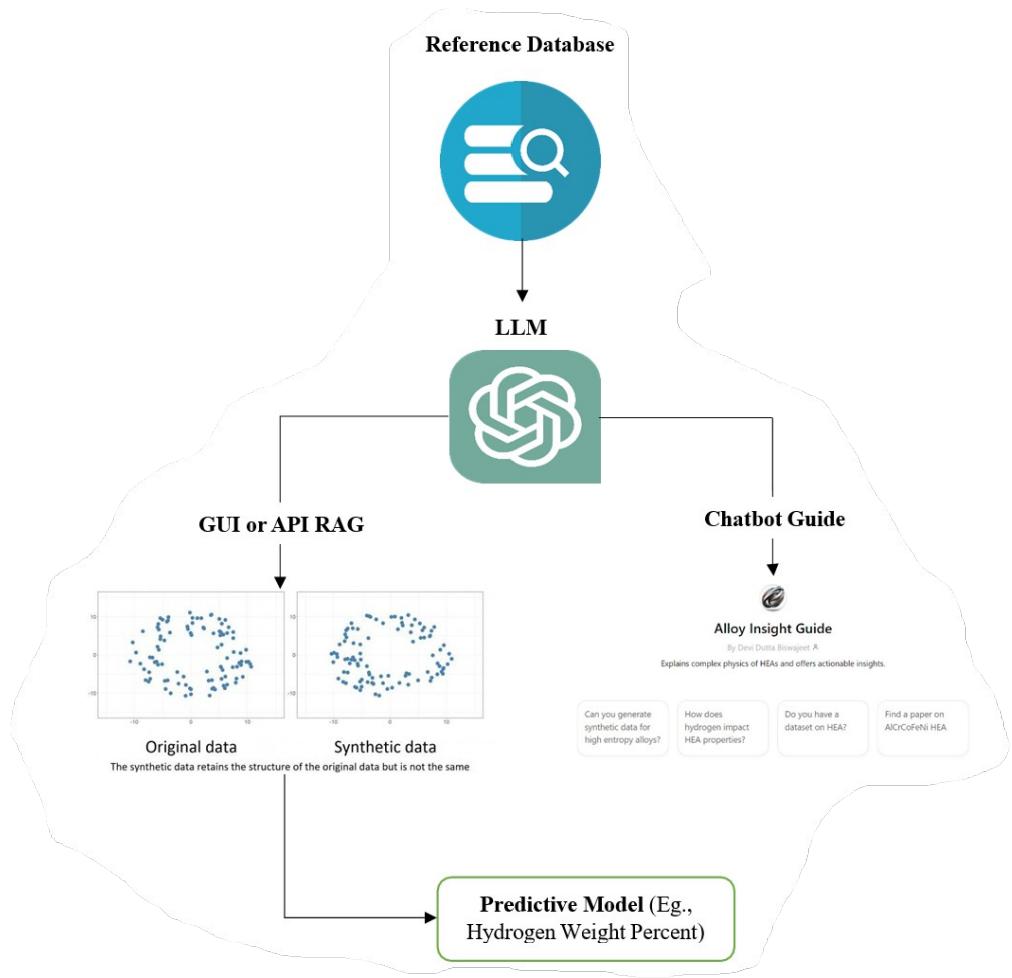


Figure 37: Insightful machine learning for HEA hydrides

- [4] Lewis, P., et al. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459-9474.
- [5] OpenAI. (2020). OpenAI GPT-3: Language models are few-shot learners. Retrieved from <https://openai.com/blog/openai-api>
- [6] OpenAI. (2023). New embedding models and API updates. Retrieved from <https://openai.com/index/new-embedding-models-and-api-updates>

# 31 Chemsense: Are large language models aligned with human chemical preference?

Authors: Marti~no R\'ios-Garc\'ia, Nawaf Alampara, Mara Schilling-Wilhelmi, Abdelrahman Ibrahim, Kevin Maik Jablonka

## 31.1 Introduction

Generative AI models are revolutionizing molecular design by learning from vast chemical datasets to create novel compounds [1]. The challenge in molecular design goes beyond just creating new structures. A key issue is ensuring the designed molecules have specific useful properties, e.g., high solubility, high synthesizability, etc. LLMs that can be conditioned on the desired property seem to be a promising solution [4]. However, current frontier models often lack an innate chemical understanding, which can lead to impractical or even dangerous molecular designs [5].

A factor that distinguishes many successful chemists is their chemical intuition. This intuition, for instance, describes the preference for certain compounds that are not grounded in knowledge that can be easily conveyed but rather by tacit knowledge accumulated over years of experience. If models could possess this chemical intuition, they would be more useful for real-world scientific applications.

In this work, we introduce ChemSense, in which we explore how well frontier models are aligned with human experts in chemistry. By aligning AI with human knowledge and preferences, one can aim to create tools that can propose feasible, safe, and desirable molecular structures, bridging the gap between theoretical capabilities and practical scientific needs. Moreover, ChemSense would help us in understanding the emergent alignment of frontier models with dataset scale and size.

## 31.2 Related work

Yang et al. (2024) investigated the ability of LLMs to predict human characteristics in their social life and showed that LLMs encountered great difficulties in predicting these. Mirza et al. (2024) benchmarked LLMs on various chemistry tasks; however, good performance in that benchmark does not guarantee alignment with human-expert intuitions. Chennakesavalu et al. (2024) aligned LLMs to generate low energy stable molecules with externally specified properties, they noticed huge improvement upon alignment compared to the base model.

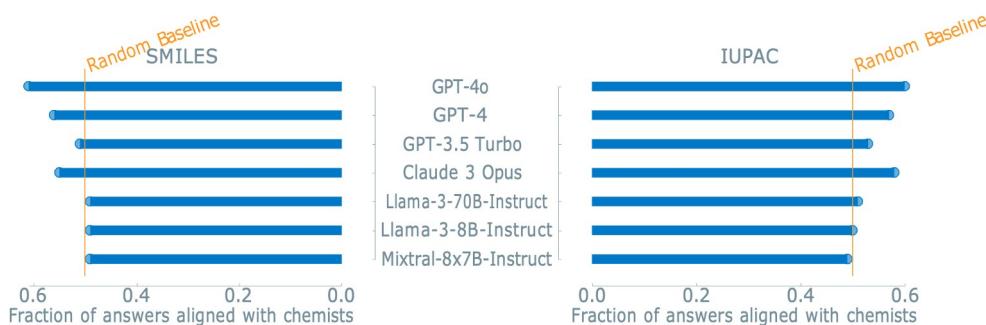


Figure 38: Comparison of the alignment of the different LLMs with the SMILES (left) and IUPAC (right) molecular representations. For both representations, note that the random baseline is at a fixed value of 0.5 since all the questions were binary and the datasets are balanced.

## 31.3 Materials and Methods

**Chemical preference dataset** The test set used during all the experimentation was constructed with 900 samples from the dataset published by Choung et al. (2023). This dataset contains the preferences of 35 Novartis medical chemists, resulting in more than 5000 question-answer pairs about organic synthesis. To

choose the 900 samples, only questions in which both molecules could be converted into the IUPAC names (using the [chemistry name resolver](#)) were selected. In that way, we ensure some homogeneity in the test dataset.

**LLM evaluation** To study the chemical preference of the actual LLMs, some of the models performing best on the ChemBench benchmark [5] were prompted with a simple instruction prompt inspired by the question that was asked to collect the original data from the scientists of Novartis. Additionally, to study how molecular representations could affect the model, each of the 900 questions was given to the model using different molecular representations.

To ensure the correct format of the answers, OpenAI models as well as Claude-3 were constrained to answer only “A” or “B” using the [Instructor](#) package. On the other hand, the Llama models and Mixtral 8Bx7 used (in an unconstrained setup) through [Groq API service](#) and instead further prompted to encourage them only to answer “A” or “B”.

## 31.4 Results

**Comparison of chemical preference of LLMs** We compare the accuracy of the preference prediction of the different models and representations (Figure 1). The accuracy of all models ranges from 49% to 61% where 50% is the accuracy one would obtain for a random choice. The GPT-4o model achieves the highest accuracy of all models and performs best with the SMILES and IUPAC representations. This might be explained by the widespread use of both of the representations and, therefore, a high occurrence in the training data of the model. We observe the same trend in the GPT-4 and GPT-3.5 Turbo predictions. For the other LLMs, representation seems to have a smaller impact on the accuracy with values that presumably are random.

## 31.5 Discussion and Conclusions

Our work shows that while some LLMs might show sparks of chemical intuition, they are mostly unaligned with expert intuitions, as their performance currently is only marginally better than random guessing. Interestingly, similar to previous work [4], we find that one obtains different performance in different molecular representations, which might be attributable to different prevalences in the training data. Our research demonstrates that preference tuning is a promising and underexplored approach to enhance models for chemists. Addressing model biases is crucial for ensuring fair and accurate predictions. By tackling these challenges, we can develop large language models (LLMs) that are both powerful and practical for chemists and materials scientists.

## References

- [1] Bilodeau, Camille et al. (2022). “Generative models for molecular discovery: Recent advances and challenges”. In: Wiley Interdisciplinary Reviews: Computational Molecular Science 12.5, e1608.
- [2] Chennakesavalu, Shriram et al. (2024). Energy Rank Alignment: Using Preference Optimization to Search Chemical Space at Scale. DOI: [10.48550/ARXIV.2405.12961](https://doi.org/10.48550/ARXIV.2405.12961) URL: <https://arxiv.org/abs/2405.12961>
- [3] Choung, Oh-Hyeon et al. (Oct. 2023). “Extracting medicinal chemistry intuition via preference machine learning”. In: Nature Communications 14.1. ISSN: 2041-1723. DOI: 10.1038/s41467-023-42242-1. URL: <http://dx.doi.org/10.1038/s41467-023-42242-1>
- [4] Jablonka, Kevin Maik et al. (Feb. 2024). “Leveraging large language models for predictive chemistry”. In: Nature Machine Intelligence 6.2, pp. 161–169. ISSN: 2522-5839. DOI: 10.1038/s42256-023-00788-1. URL: <http://dx.doi.org/10.1038/s42256-023-00788-1>
- [5] Mirza, Adrian et al. (2024). “Are large language models superhuman chemists?” In: arXiv preprint. DOI: 10.48550/arXiv.2404.01475. arXiv: 2404.01475 [cs.LG].

[6] Yang, Kaiqi et al. (2024). Are Large Language Models (LLMs) Good Social Predictors? DOI: 10.48550/ARXIV.2402.12620. URL: <https://arxiv.org/abs/2402.12620>

## 32 GlossaGen

Authors: *Magdalena Lederbauer, Dieter Plessers, Philippe Schwaller*

Academic articles, particularly reviews, and grant proposals would greatly benefit from a glossary explaining complex jargon and terminology. However, the manual creation of such glossaries is a time-consuming and repetitive task. To address this challenge, we developed GlossaGen, an innovative framework that leverages large language models to automatically generate glossaries from PDF or TeX files, streamlining the process for academics. The generated glossary is not only a list of terms and definitions but also visualized as a knowledge graph, illustrating the intricate relationships between various technical concepts (see Figure 39).

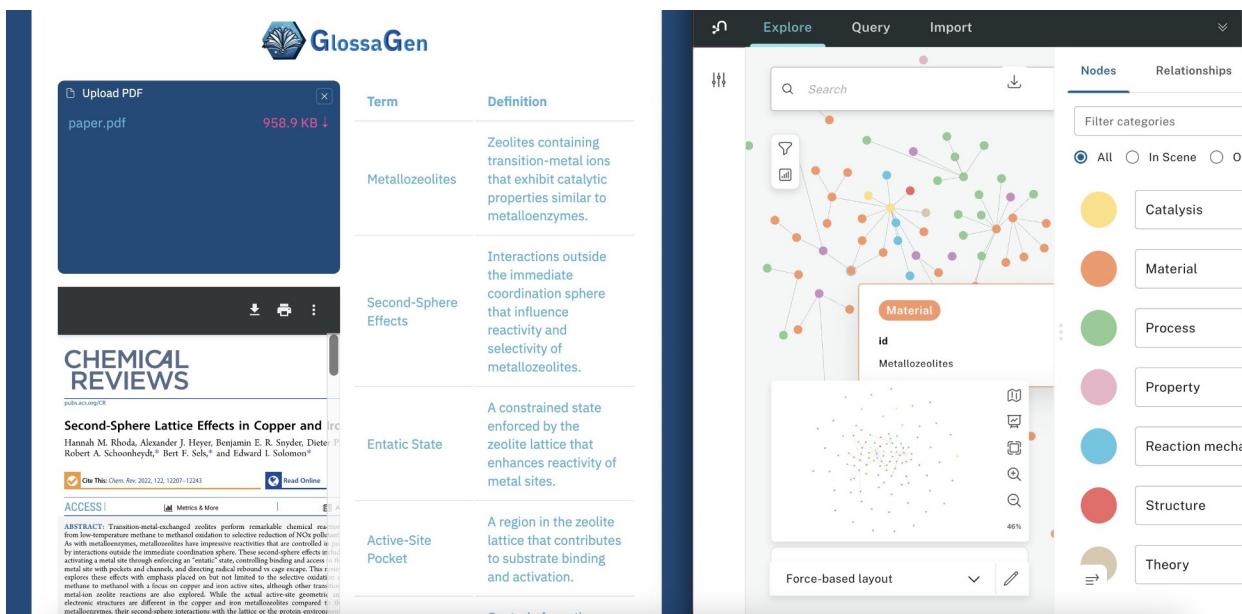


Figure 39: Overview of (left) the graphical user interface (GUI) prototype and (right) the generated Neo4J knowledge graph (right). Our results demonstrate that LLMs can greatly accelerate glossary creation, increasing the likelihood that authors will include a helpful glossary without the need for tedious manual effort. Additionally, an analysis of our test case by a zeolite domain expert revealed that LLMs produce good results, with about 70% - 80% of explanations requiring little to no manual changes.

The project's codebase was developed as a Python package on GitHub using a template [1] and DSPy [2] as an LLM framework. This modular approach facilitates seamless collaboration and easy incorporation of new features.

To overcome the limitations of LLMs in directly processing PDFs, a prevalent format for scientific publications, we implemented a pre-processing step that converts papers into manageable text sequences. This step involves extracting textual information using PyMuPDF [3], automatically obtaining the title and DOI, and chunking the text into smaller sections. This chunking preserves context and makes it easier for the models to handle the input.

We used GPT-3.5-Turbo [4] and GPT-4-Turbo [5] to extract scientific terms and their definitions from the text chunks. Employing Typed Predictors [6] and Chain-of-Thought prompting [7] ensures the outputs are well-structured and contextually accurate, guiding the model to produce precise definitions through a simulated reasoning process. Post-processing involved identifying and removing duplicate entries, ensuring each term appears only once in the final glossary. Figure 40 shows details about the GlossaryGenerator class that was used to process documents into the corresponding glossaries. We selected a review article on zeolites [8] (shown in Figure 39) as a test publication to manually tune and evaluate the pipeline's output.

The obtained glossary is transformed into an ontology that defines nodes and relationships for the knowledge graph. For instance, relationships like 'MATERIAL – exhibits → PROPERTY' illustrate how different terms are interconnected. The knowledge graph is constructed using the library Neo4J [9] and Graph

```

class GlossaryGenerator:
    """
    A class that generates a glossary based on a research document.

    Attributes
    -----
    research_doc (ResearchDoc): The research document to generate the glossary from.
    glossary_predictor (dspy.Predict): The predictor used to generate the glossary.
    chunk_size (int): The size of the chunks to split the research document into.
    reranker (dspy.TypedChainOfThought): The reranker used to filter important terms.

    Methods
    -----
    __init__(self, research_doc: ResearchDoc, chunk_size: int = 20000):
        Initialize a GlossaryGenerator object.

    normalize_term(self, term: str) -> str:
        Normalize a term by converting it to lowercase and removing common plural endings.

    deduplicate_entries(self, glossary: list[TerminusTechnicus]) -> list[TerminusTechnicus]:
        Deduplicate the glossary entries by considering plurals and similar-sounding terms.

    format_nicely(self, glossary: list[TerminusTechnicus]) -> str:
        Format the glossary nicely.

    generate_glossary_from_doc(self) -> pd.DataFrame:
        Generate the glossary based on the research document.

    """

```

Figure 40: Overview of the GlossaryGenerator class, responsible for processing text chunks and extracting relevant, correctly formatted terms and definitions.

Marker [10] on the processed text chunks. We developed a user-friendly front-end interface with Gradio [11], as shown in Figure 39. This interface allows users to interact with the glossary, making it easier to navigate and customize the information.

The quick prototyping provided us with several ideas for future work. We can improve the glossary output by fine-tuning the LLM, incorporating retrieval-augmented generation, and parsing article images. Additionally, the user experience can be enhanced by allowing users to input specific terms for glossary explanations as a backup when the LLM omits certain terms. Integration with LaTeX would broaden usability, envisioning commands like `\glossarysimilar{}` to `\bibliography{}`. We also consider connecting the knowledge graph directly to the user interface and enhance its ontology creation feature. Overall, this rapidly developed prototype, with numerous future possibilities, demonstrates the potential of LLMs to assist researchers in their scientific outreach.

## References

- [1] Copier template: Available at <https://github.com/copier-org/copier>
- [2] DSPy: Khattab, O.; Singhvi, A.; Maheshwari, P.; Zhang, Z.; Santhanam, K.; Vardhamanan, S.; Haq, S.; Sharma, A.; Joshi, T. T.; Moazam, H.; Miller, H.; Zaharia, M.; Potts, C. DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines. preprint [arXiv:2310.03714](https://arxiv.org/abs/2310.03714). 2023.
- [3] PyMuPDF: Available at <https://github.com/pymupdf/PyMuPDF>
- [4] GPT-3.5-Turbo: OpenAI. Available at <https://platform.openai.com/docs/models/gpt-3-5-turbo>

- [5] GPT-4-Turbo: OpenAI. Available at <https://platform.openai.com/docs/models/gpt-4-turbo-and-gpt-4>
- [6] DSPy Typed Predictors: Documentation at [https://dspy.ai/learn/8-typed\\_predictors](https://dspy.ai/learn/8-typed_predictors)
- [7] Chain-of-Thought prompting: Wei, J.; Wang X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; Zhou, D. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. preprint arXiv: 2201.11903. 2023.
- [8] Test review article on zeolites: Rhoda, H. M.; Heyer, A. J.; Snyder, B. E. R.; Plessers, D.; Bols, M. L.; Schoonheydt, R. A.; Sels, B. F.; Solomon, E. I. Second-Sphere Lattice Effects in Copper and Iron Zeolite Catalysis. *Chem. Rev.* 2022, 122, 12207–12243.
- [9] Neo4J: Documentation at <https://neo4j.com/>
- [10] Graph Maker: Available at [https://github.com/rahulnyk/graph\\_maker](https://github.com/rahulnyk/graph_maker)
- [11] Gradio: Available at <https://github.com/gradio-app/gradio>