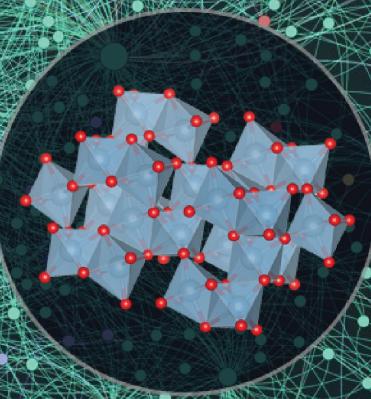
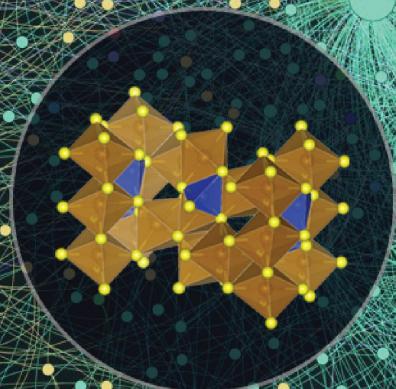
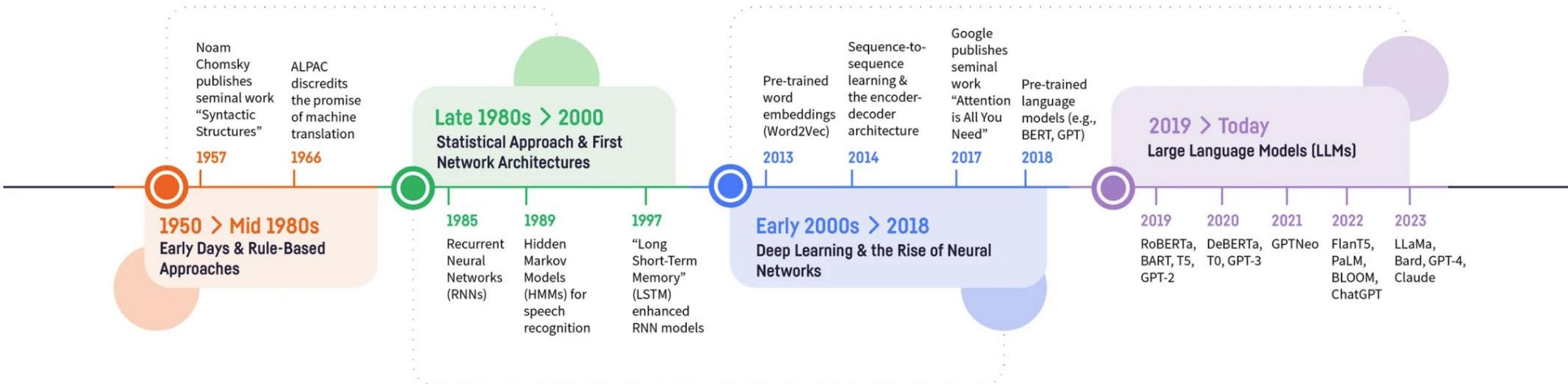


Large Language Models in Materials



Artificial intelligence has been applied to language from the beginning



So what is NLP and language models?

Natural Language Processing (NLP) is a branch of AI that gives machines the ability to read, understand, and derive meaning from human languages

Language models are algorithms that predict the likelihood of a sequence of words

Large Language Models (LLMs) are just language models deployed at a comprehensive scale

Let's examine the math of pre neural network NLP

Bag of Words: Representation of text as the bag (multiset) of its words, disregarding grammar and even word order but keeping multiplicity.

- Akin to one-hot-encoding you create a vector where each vector dimension is a unique word in the series and the value of that dimension is the word frequency

Term Frequency – Inverse Document Frequency: TF-IDF is a numerical statistic intended to reflect how important a word is to a document in a collection or corpus.

$$\left(\frac{\# \text{ of times term } t \text{ appears in document}}{\text{total number of terms in document}} \right) * \log \left(\frac{\text{total number of documents}}{\text{number of documents with term } t} \right)$$

Let's examine the math of pre neural network NLP

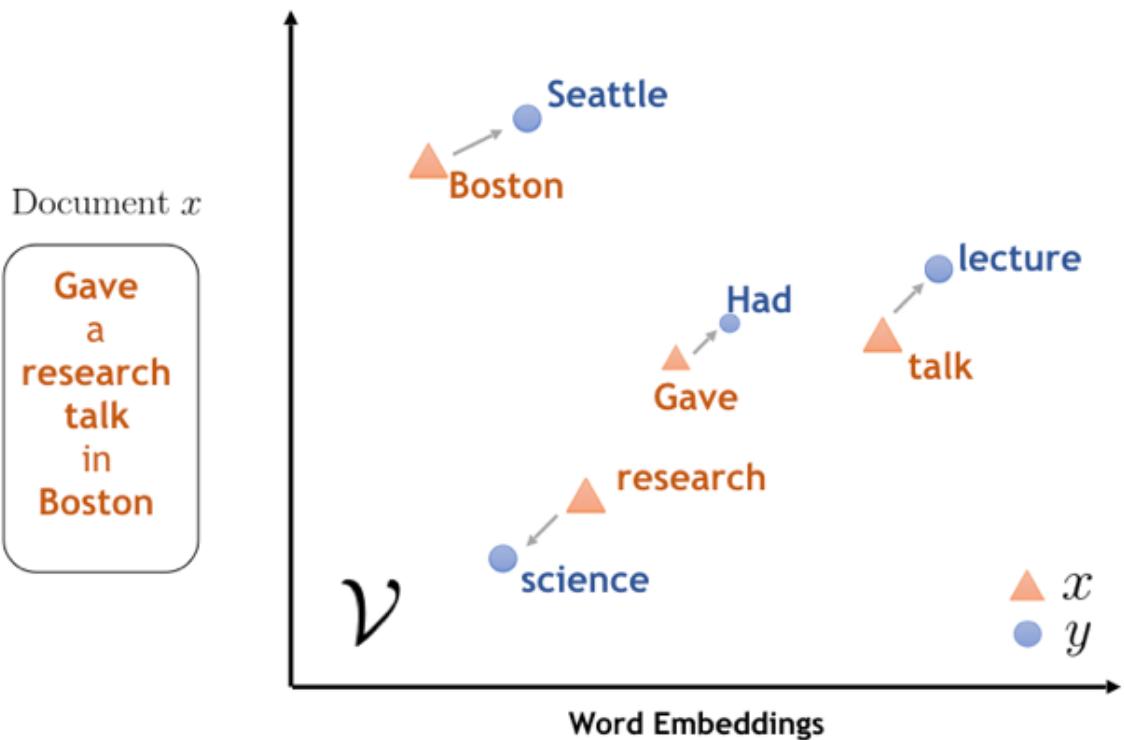
N-gram models: Probabilistic models used to predict the next item in a sequence as a function of the n-1 preceding items.

- $P(w_n | w_{n-1}, \dots, w_1)$
- We often estimate this as the max likelihood estimation from a corpus of text

Latent Semantic Analysis: analyze relationships between a set of documents and the terms they contain by producing a set of concepts related to the documents and terms.

- First create a term-document matrix
- Decompose into U, Σ, V^T
- Σ shows ranked order of values. We keep top k rows to reduce dimensionality while preserving patterns.

Word embeddings have evolved in the neural network era



Google Code Archive

Projects Search About

Project word2vec

Source

Issues

Wikis

Downloads

Tool for computing continuous distributed representations of words.

Introduction

This tool provides an efficient implementation of the continuous bag-of-words and skip-gram architectures for computing vector representations of words. These representations can be subsequently used in many natural language processing applications and for further research.

Quick start

- Download the code: svn checkout <http://word2vec.googlecode.com/svn/trunk/>
- Run 'make' to compile word2vec tool
- Run the demo scripts: `./demo-word.sh` and `./demo-phrases.sh`
- For questions about the toolkit, see <http://groups.google.com/group/word2vec-toolkit>

Project Information

The project was created on Jul 29, 2013.

- License: Apache License 2.0
- 945 stars
- svn-based source control

Labels:

- NeuralNetwork MachineLearning
NaturalLanguageProcessing WordVectors
Google

GloVe: Global Vectors for Word Representation

Jeffrey Pennington, Richard Socher, Christopher D. Manning

Introduction

GloVe is an unsupervised learning algorithm for obtaining vector representations for words. Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space.

Getting started (Code download)

- Download the latest [latest code](#) (licensed under the [Apache License, Version 2.0](#)). Look for "Clone or download"
- Unpack the files: `unzip master.zip`
- Compile the source: `cd GloVe-master & make`
- Run the demo script: `./demo.sh`
- Consult the included README for further usage details, or ask a [question](#)

Download pre-trained word vectors

- Pre-trained word vectors. This data is made available under the [Public Domain Dedication and License v1.0](#) whose full text can be found at: <http://www.opendatacommons.org/licenses/pddl/1.0/>
 - Wikipedia 2014 + Gigaword 5 (6B tokens, 400M vocab, uncased, 50d, 100d, 200d, & 300d vectors, 822 MB download): [glove.6B.zip](#)
 - Common Crawl (12B tokens, 19M vocab, uncased, 300d vectors, 1.75 GB download): [glove.6B.100d.zip](#)
 - Common Crawl (34GB tokens, 2.2M vocab, cased, 300d vectors, 2.03 GB download): [glove.840B.300d.zip](#)
 - Twitter (2B tweets, 2B tokens, 1.2M vocab, uncased, 25d, 50d, 100d, & 200d vectors, 1.42 GB download): [glove.twitter.27B.zip](#)
- Ruby `script` for preprocessing Twitter data

Citing GloVe

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. [GloVe: Global Vectors for Word Representation](#) ([pdf](#) | [bib](#))

Let's describe the math behind Word2Vec

Word2Vec: two versions Continuous Bag of Words (CBOW) and Skip-Gram

- Skip-gram: given a target word, w_I , the model predicts surrounding context words, w_O
- Softmax to calculate conditional probability

$$P(w_O|w_I) = \frac{\exp({v'_{w_O}}^T v_{w_I})}{\sum_{w=1}^W \exp({v'_{w}}^T v_{w_I})}$$

- Where v are vectors representations of words, W is vocab size
- Maximize this probability for all correct word-context pairs in corpus
- CBOW: similar to Skip-gram, but inverts the role of target and context words

Let's describe the math behind GloVe

GloVe: combines matrix factorization and local context window methods. It constructs an explicit word-context (or word co-occurrence) matrix by counting how frequently a word appears in a context.

- Co-occurrence matrix X_{ij} for word i in context of word j.
 - Context just means existing within some window of words surrounding i
- Objective function: train embeddings by directly relating word vectors to the probabilities that two words appear together. Minimize the difference between the dot product of the embeddings of two words and the log of their co-occurrence probability:

$$J = \sum_{i,j=1}^V f(X_{ij}) (w_i^T w_j + b_i + b_j - \log X_{ij})^2$$

NLP uses the same sequential data algorithms we've already learned

Recurrent neural networks (RNNs)

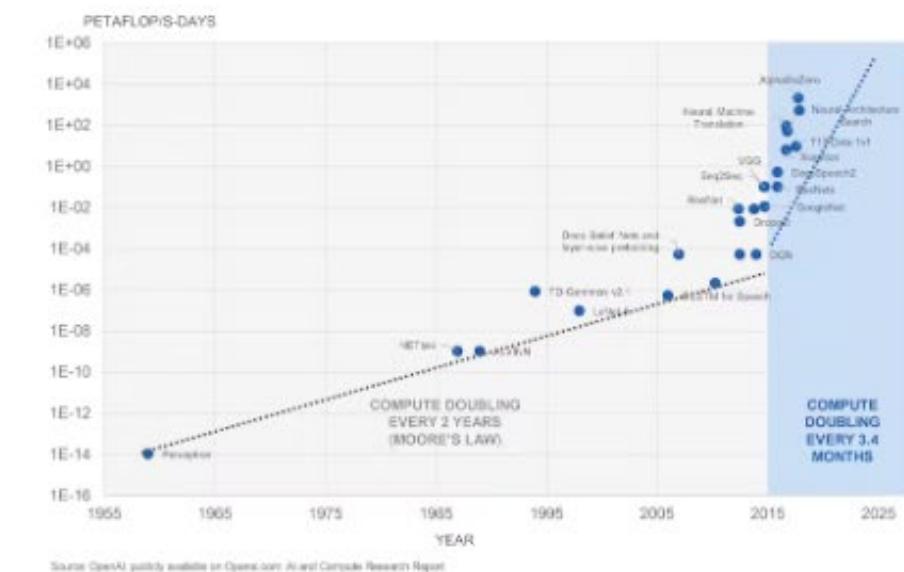
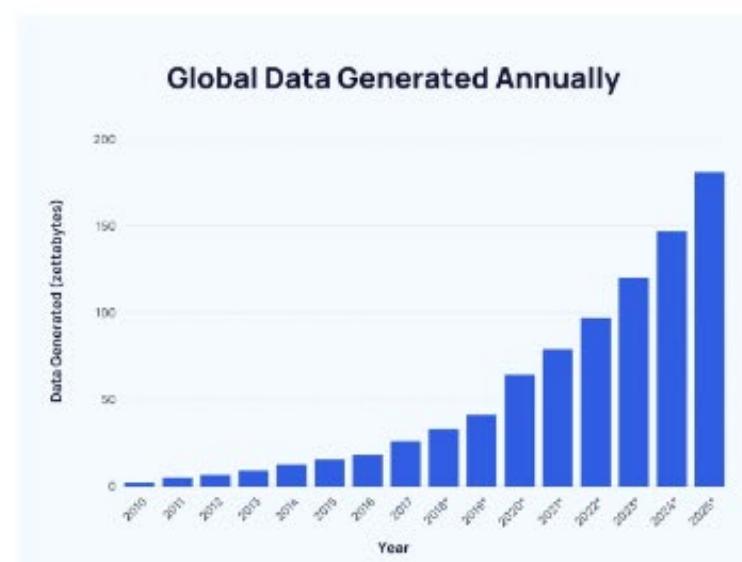
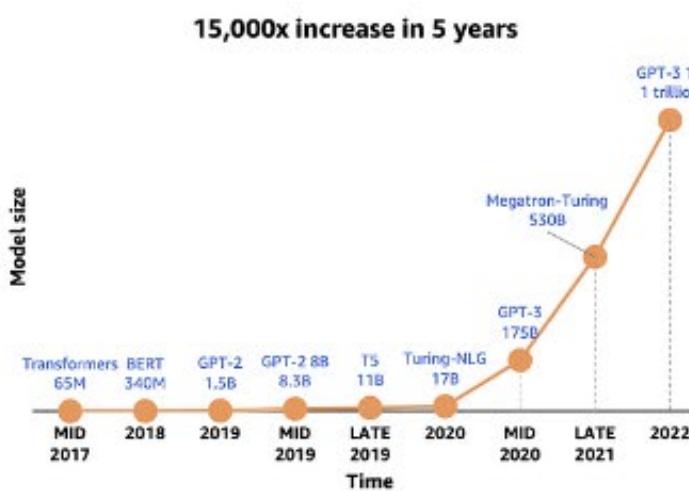
Long Short-Term Memory (LSTM)

Gated Recurrent Unit (GRU)

Transformers

LLMs are very big
versions of these,
mostly Transformers

Why are LLMs so large?



Kaplan et al., Scaling Laws for Neural Language Models, 2020

 Yann LeCun  
@ylecun

On the highway towards Human-Level AI, Large Language Model is an off-ramp.

2:39 AM · Feb 4, 2023 · 1.5M Views

274 484 2.9K 423

Milestones in LLM development

GPT (Generative Pre-trained Transformer, OpenAI)

Overview: A series of models that have set new standards for generative text tasks. GPT models are trained on vast amounts of text data and can generate coherent, contextually relevant text based on a given prompt. Each version of GPT has significantly increased in size and capability, with GPT-3 being one of the most widely known.

Capabilities: Text generation, translation, summarization, question-answering, and more.

BERT (Bidirectional Encoder Representations from Transformers, Google)

Overview: BERT marked a significant shift in how contextual information is processed by LLMs. It uses a mechanism that allows it to consider the full context of a word by looking at the words that come before and after it in a sentence, making it highly effective for understanding the nuances of language.

Capabilities: Exceptional performance on tasks like natural language understanding, sentiment analysis, and entity recognition.

T5 (Text-to-Text Transfer Transformer, Google)

Overview: T5 is designed to convert all NLP problems into a unified text-to-text format, where both input and output are always text strings. This approach simplifies the training process and allows the model to be applied across a diverse range of tasks without task-specific modifications.

Capabilities: Summarization, question answering, text classification, and translation.

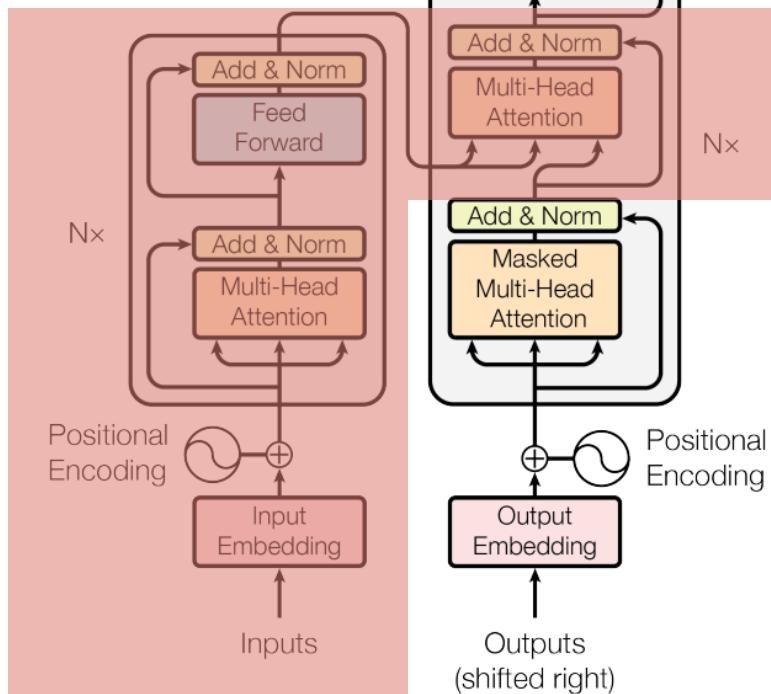
XLNet (Google Brain and Carnegie Mellon University)

Overview: XLNet introduces a novel training strategy that outperforms BERT on several NLP benchmarks. It uses a permutation-based training method that allows it to capture bidirectional context and outperform BERT in tasks requiring a deep understanding of context.

Capabilities: State-of-the-art performance on tasks like text classification, question answering, and more.

GPTs actually only use part of the transformer architecture

Not present
in GPT

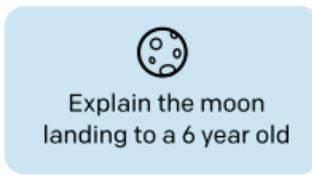


LLM Training

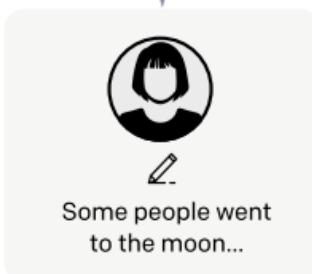
1. Tokenize input
2. Pass input embeddings through model
3. Produce output
4. Calculate error
5. Backpropagation to adjust weights

We can get **alignment** via reinforcement learning from human feedback

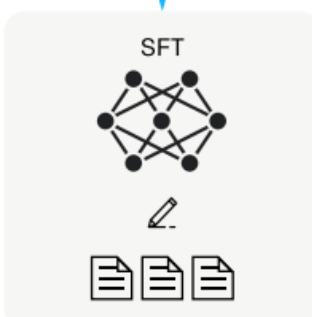
A prompt is sampled from our prompt dataset.



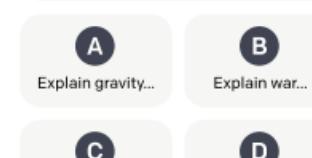
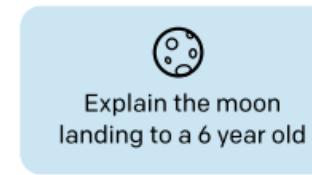
A labeler demonstrates the desired output behavior.



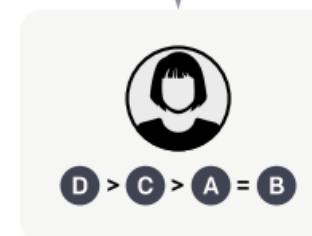
This data is used to fine-tune GPT-3 with supervised learning.



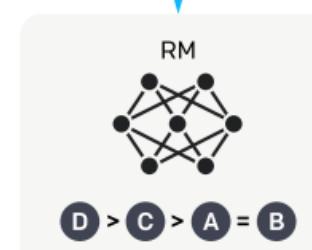
A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.

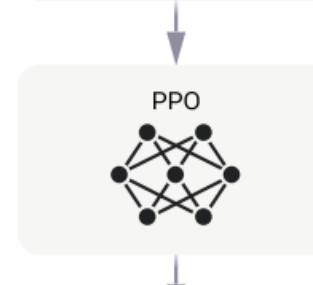


This data is used to train our reward model.



A new prompt is sampled from the dataset.

The policy generates an output.



Once upon a time...



The reward model calculates a reward for the output.

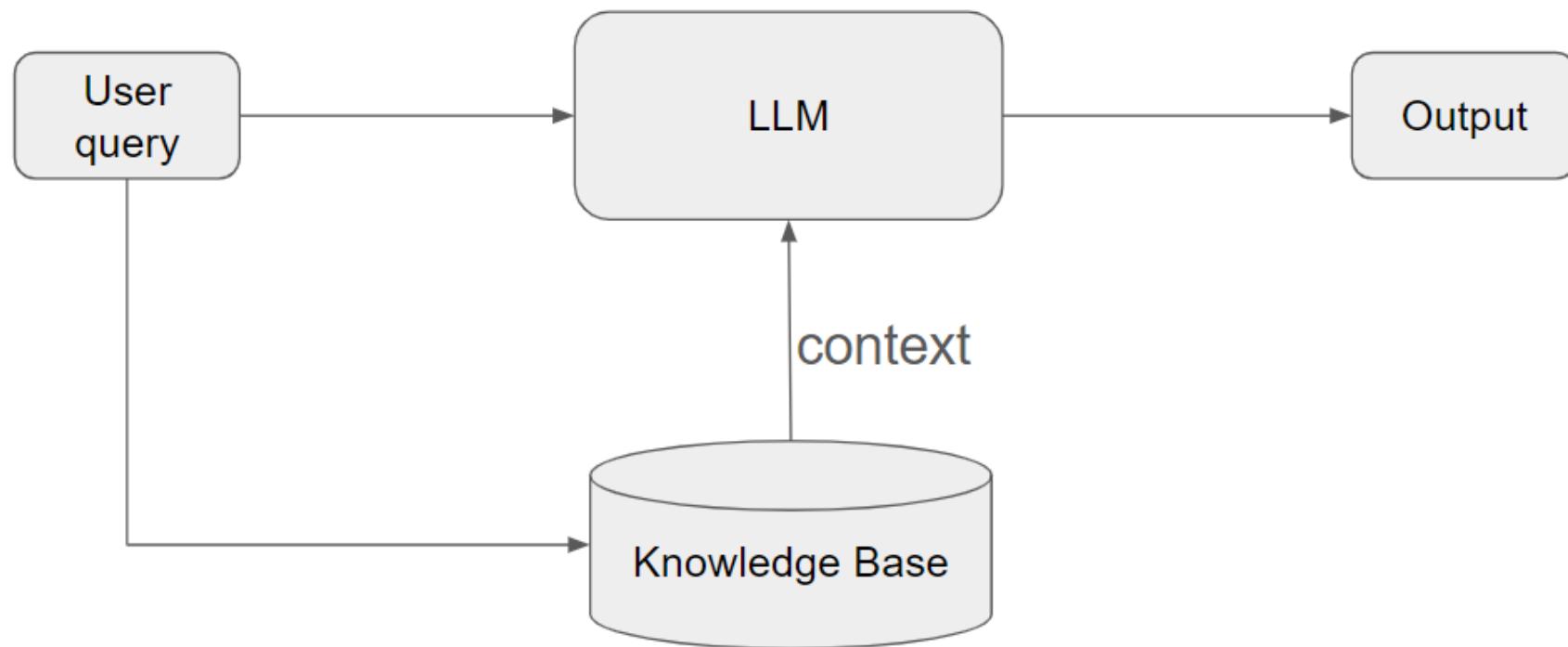
The reward is used to update the policy using PPO.

r_k



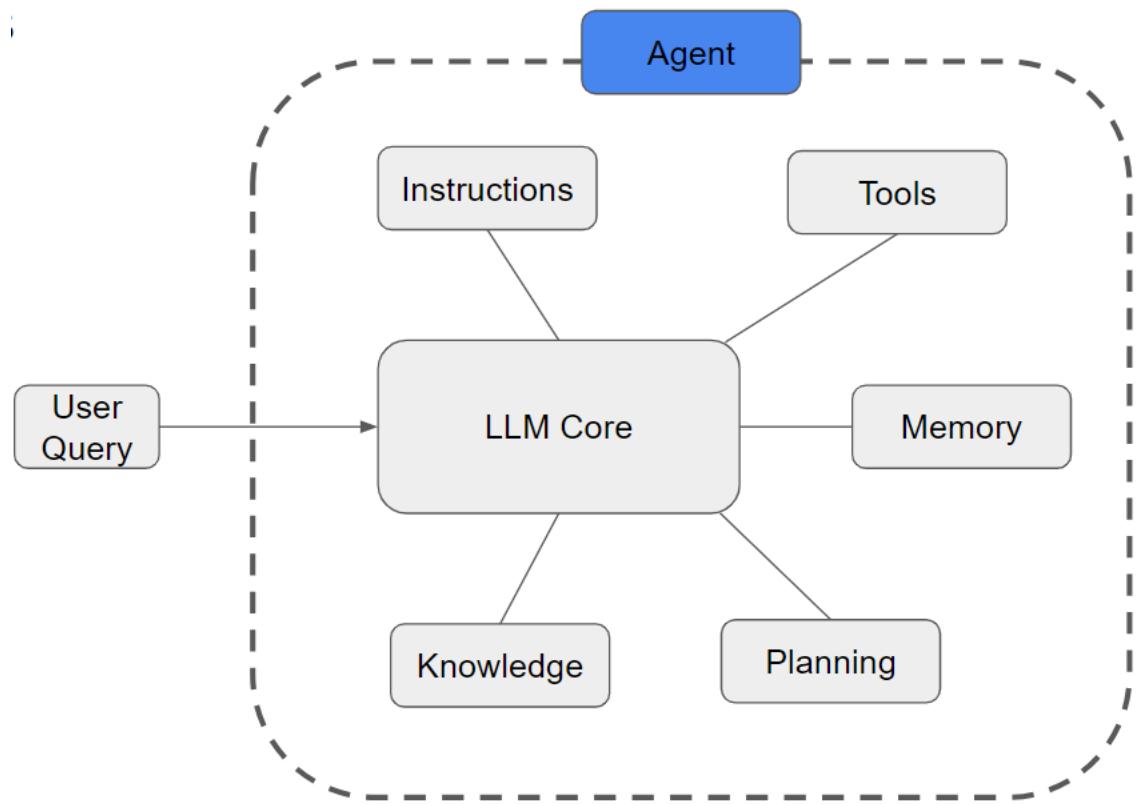
Retrieval Augmented Generation (RAG) offers solutions to LLM problems

LLMs have token size limitations and training data cutoff date.



Perplexity.ai is a really good example of this

Agents offer another application of LLMs



Screenshot of the MetaGPT GitHub repository (Public). The repository has 832 watchers, 4.4k forks, and 37.7k stars. It contains 21 branches and 20 tags. The 'Code' tab is selected, showing a list of recent commits:

Commit	Message	Date
6450a09 · yesterday	Merge pull request #1114 from better629/feat_memory	3,729 Commits
.devcontainer	add proper space	3 months ago
.github	simplify provider ut code	last month
config	refactor: timeout 0	last week
docs	Update README_JA.md	2 weeks ago
examples	rm unnecessary comment	yesterday
metagpt	Merge branch 'main' into feat_memory	yesterday
tests	Merge branch 'main' into feat_memory	yesterday
.dockerrcignore	remove duplicate string	3 months ago
.gitattributes	add index factory and chromadb	3 weeks ago

About

- The Multi-Agent Framework: First AI Software Company, Towards Natural Language Programming
- [deepwisdom.ai/](#)
- agent multi-agent gpt hacktoberfest llm metagpt
- Readme
- MIT license
- Security policy
- Activity
- 37.7k stars
- 832 watching
- 4.4k forks

Report repository

Now let's discuss how LLMs are being used in materials science

Data extraction

Property prediction

Generative models

Automation and novel interfaces

Education

Digital
Discovery

PERSPECTIVE



[View Article Online](#)
[View Journal](#) | [View Issue](#)

Check for updates

Cite this: *Digital Discovery*, 2023, 2,
1233

Open Access Article. Published on 08 August 2023. Downloaded on 3/28/2024 5:24:13 AM.

This article is licensed under a Creative Commons Attribution 3.0 Unported Licence.

14 examples of how LLMs can transform materials science and chemistry: a reflection on a large language model hackathon[†]

Kevin Maik Jablonka,^b* Qianxiang Ai,^b* Alexander Al-Feghali,^b* Shruti Badhwar,^{a,d} Joshua D. Bocarsly,^b* Andres M. Bran,^b*^{f,g} Stefan Bringuiel,^b*ⁱ L. Catherine Brinson,^b* Kamal Choudhary,^b* Define Cirić,^b* Sam Cox,^b* Wibe A. de Jong,^b* Matthew L. Evans,^b*^{mn} Nicola Gastello,^b* Jerome Genzling,^b*^c María Victoria Gil,^b*^o Ankur K. Gupta,^b* Zhi Hong,^b*^d Alishba Imran,^b* Sabina Kruschwitz,^b*^r Anne Labarre,^b* Jakub Lála,^b*^s Tao Liu,^b*^t Steven Ma,^b*^c Sauradeep Majumdar,^b*^a Garrett W. Merz,^b*^t Nicolas Moitessier,^b*^c Elias Moubarak,^b*^a Beatriz Mourão,^b*^t Brenden Pelkie,^b*^t Michael Pieler,^b*^s Mayk Caldas Ramos,^b*^t Bojana Ranković,^b*^{t,g} Samuel G. Rodrigues,^b*^s Jacob N. Sanders,^b*^t Philippe Schwaller,^b*^{t,g} Marcus Schwarting,^b*^y Jiale Shi,^b*^t Berend Smit,^b*^a Ben E. Smith,^b*^e Joren Van Herck,^b*^a Christoph Völker,^b*^t Logan Ward,^b*^{t,z} Sean Warren,^b*^t Benjamin Weiser,^b*^t Sylvester Zhang,^b*^c Xiaoqi Zhang,^b*^a Ghézal Ahmad Zia,^b*^t Aristana Scourtas,^b*^{aa} K. J. Schmidt,^b* Ian Foster,^b*^k Andrew D. White,^b*^k and Ben Blaizsik,^b*^{aa}

Received 12th June 2023

Accepted 8th August 2023

DOI: 10.1039/d3dd00113j

rsc.li/digitaldiscovery

Large-language models (LLMs) such as GPT-4 caught the interest of many scientists. Recent studies suggested that these models could be useful in chemistry and materials science. To explore these possibilities, we organized a hackathon. This article chronicles the projects built as part of this hackathon. Participants employed LLMs for various applications, including predicting properties of

^aLaboratory of Molecular Simulation (LSMO), Institut des Sciences et Ingénierie Chimiques, Ecole Polytechnique Fédérale de Lausanne (EPFL), Sion, Valais, Switzerland. E-mail: mail@jablonka.com

^bDepartment of Chemical Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

^cDepartment of Chemistry, McGill University, Montreal, Quebec, Canada

^dReinhardt Inc., USA

^eYusaf Hamied Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge, CB3 1EW, UK

^fLaboratory of Artificial Chemical Intelligence (LACI), Institut des Sciences et Ingénierie Chimiques, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

^gNational Centre of Competence in Research (NCCR) Catalysis, Ecole Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland

^hIndependent Researcher, San Diego, CA, USA

ⁱMechanical Engineering and Materials Science, Duke University, USA

^jMaterial Measurement Laboratory, National Institute of Standards and Technology, Maryland, 20899, USA

^kDepartment of Chemical Engineering, University of Rochester, USA

^lApplied Mathematics and Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

^mInstitut de la Matière Condensée et des Nanosciences (IMCN), UCLouvain, Chemin des Étoiles 8, Louvain-la-Neuve, 1348, Belgium

ⁿMatgenix SRL, 185 Rue Armand Bury, 6534 Gozée, Belgium

^oInstituto de Ciencia y Tecnología del Carbono (INCAR), CSIC, Francisco Pintado Fe 26, 33011 Oviedo, Spain

^pDepartment of Computer Science, University of Chicago, Chicago, Illinois 60637, USA

^qComputer Science, University of California, Berkeley, CA 94704, USA

^rBundesanstalt für Materialforschung und -prüfung, Unter den Eichen 87, 12205 Berlin, Germany

^sFrancis Crick Institute, 1 Midland Rd, London NW1 1AT, UK

^tAmerican Family Insurance Data Science Institute, University of Wisconsin-Madison, Madison, WI 53706, USA

^uDepartment of Chemical Engineering, University of Washington, Seattle, WA 98105, USA

^vOpenBioMol.org, UK

^wStabilityAI, UK

^xDepartment of Chemistry and Biochemistry, University of California, Los Angeles, CA 90095, USA

^yDepartment of Computer Science, University of Chicago, Chicago, IL 606490, USA

^zData Science and Learning Division, Argonne National Lab, USA

^{aa}Globus, University of Chicago, Data Science and Learning Division, Argonne National Lab, USA. E-mail: blaiszik@uchicago.edu

[†]Department of Computer Science, University of Chicago, Data Science and Learning Division, Argonne National Lab, USA

[‡]Electronic supplementary information (ESI) available. See DOI:

<https://doi.org/10.1039/d3dd00113j>

^{*}These authors contributed equally.

Semantic text mining approaches have been around for a while

Hawizy et al. *Journal of Cheminformatics* 2011, 3:17
http://www.jcheminf.com/content/3/1/17



RESEARCH ARTICLE

ChemicalTagger: A tool for semantic text-mining in chemistry

Lezan Hawizy^{1*}, David M Jessop^{1*}, Nico Adams² and Peter Murray-Rust¹

Abstract

Background: The primary method for scientific communication is in the form of published scientific theses which use natural language combined with domain-specific terminology. As such, they contain unstructured text. Given the usefulness of data extraction from unstructured literature, we aim to show what can be achieved for the discipline of chemistry. The highly formulaic style of writing most chemists share their contributions well suited to high-throughput Natural Language Processing (NLP) approaches.

Results: We have developed the ChemicalTagger parser as a medium-depth, phrase-based semantic parser of the language of chemical experiments. Tagging is based on a modular architecture and uses a core OSCAR, domain-specific regex and English taggers to identify parts-of-speech. The ANTLR grammar is structured into tree-based phrases. Using a metric that allows for overlapping annotations, we achieve annotator agreements of 88.9% for phrase recognition and 91.9% for phrase-type identification.

Conclusions: It is possible parse to chemical experimental text using rule-based techniques in conjunction with a core grammar parser. ChemicalTagger has been deployed for over 10,000 patents and has identified from their linguistic context with >99.5% precision.

Background

In many scientific disciplines, the primary method of communicating scientific results is in the form of a scientific paper or thesis which uses free flowing natural language combined with domain-specific terminology and numeric phrases. As such, they contain unstructured data, which is not identifiable by machines and not easily re-usable. Information providers have built businesses around the manual abstraction of unstructured data in the literature by human domain experts. Apart from the considerable labour cost and delay after the original publication, human abstraction is also a considerable source of error and data corruption.

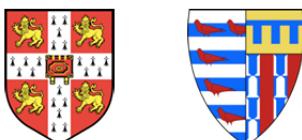
A typical synthesis procedure taken from the organic chemistry literature, reads as follows: [1]

5-Cyclobutyl-2,3-dihydro-[1H]-2-benzothiophene. Potassium carbonate (0.63 g, 4.6 mmol) and thiophenol (0.19 g, 1.69 mmol) were added to N,N-dimethylformamide (50.050 ml) in N,N-dimethylformamide (33 °C). The mixture was stirred at room temperature and the mixture was stirred. Deionised water (50 cm³) was added and the aqueous phase was extracted with ethyl acetate (3 × 50 cm³). The organic extracts were dried over magnesium sulfate and reduced pressure to give a yellow oil (0.259 g, 1.302 mmol).

The example shown here shows highly formulaic language, which occurs in many documents and is not just restricted to chemistry, as

- **Semi-structured documents:** Usual typographic conventions such as new text, rather than formal markup.
- **Domain-specific entities:** Entities from different scientific domains.

Daniel Mark Lowe
Pembroke College



This dissertation is submitted for the degree of Doctor of Philosophy

June 2012

* Correspondence: mh29@cam.ac.uk
Unilever Centre for Molecular Science Informatics, Department of Chemistry, Lensfield Road, Cambridge, CB2 1EW, UK
Full list of author information is available at the end of the article

© 2011 Hawizy et al. Licensee Chemistry Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

© 2011 Hawizy et al. Licensee Chemistry Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

DEPARTMENT OF CHEMISTRY

Jessop et al. *Journal of Cheminformatics* 2011, 3:41
http://www.jcheminf.com/content/3/1/41



SOFTWARE

OSCAR4: a flexible architecture for cheminformatics

David M Jessop, Sam E Adams, Egon L Willighagen, Lezan Hawizy and Peter Murray-Rust*

Abstract

The Open-Source Chemistry Analysis Routines (OSCAR) software, a toolkit for the recognition of chemical structures in chemistry publications, has been developed since 2002. Recent work has resulted in the core OSCAR functionality and its release as the OSCAR4 library. This library features a modular API reduction of surface coupling) that permits client programmers to easily incorporate it into existing systems. OSCAR4 offers a domain-independent architecture upon which chemistry specific text-mining tools and its development and usage are discussed.

Introduction

In keeping with the historical and methodological aspects of this special issue, we recount the history and motivation of OSCAR.

A large amount of factual data in chemistry and neighbouring disciplines is published in the form of text and components within text rather than as structured semantic information. If we can discover and extract this information, the textual literature becomes an enormous additional chemical resource. As an example, we estimate that about 10 million chemical syntheses per year are published in the public literature (articles, patents, theses) and the conventional method is a natural language narrative (most commonly in English). It is extremely tedious and error-prone to extract information from this narrative manually, and for this reason many chemical abstracting services limit their scope and also frequently lag behind the current publication list.

The discipline of text-mining has now reached a state where much natural language in textual form can be analysed rapidly and with high precision and recall. Methodologies applied to the problem of chemical named entity recognition include dictionary- and rule-based methods, as well as machine learning and hybrid approaches [1–11]. We have been working in this area for approximately 10 years and the OSCAR4 software, together with OPSIN (the Open Parser for Systematic IUPAC Nomenclature) [12,13] and ChemicalTagger

* Correspondence: pm28@cam.ac.uk
Unilever Centre for Molecular Science Informatics, Department of Chemistry, Lensfield Road, Cambridge CB2 1EW, UK
© 2011 Jessop et al. Licensee Chemistry Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

© 2011 Jessop et al. Licensee Chemistry Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

© 2011 Jessop et al. Licensee Chemistry Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

© 2011 Jessop et al. Licensee Chemistry Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

© 2011 Jessop et al. Licensee Chemistry Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

© 2011 Jessop et al. Licensee Chemistry Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

© 2011 Jessop et al. Licensee Chemistry Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

© 2011 Jessop et al. Licensee Chemistry Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

© 2011 Jessop et al. Licensee Chemistry Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

© 2011 Jessop et al. Licensee Chemistry Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

© 2011 Jessop et al. Licensee Chemistry Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

© 2011 Jessop et al. Licensee Chemistry Central Ltd. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

JOURNAL OF
CHEMICAL INFORMATION
AND MODELING

Article

pubs.acs.org/jcim

ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature

Matthew C. Swain and Jacqueline M. Cole*

Cavendish Laboratory, University of Cambridge, J. J. Thomson Avenue, Cambridge, CB3 0HE, U.K.

ABSTRACT: The emergence of “big data” initiatives has led

to the need for tools that can automatically extract valuable chemical information from large volumes of unstructured data, such as the scientific literature. Since chemical information can be present in figures, tables, and textual paragraphs, successful information extraction often depends on the ability to interpret all of these domains simultaneously. We present a complete toolkit for the automated extraction of chemical entities and their associated properties, measurements, and relationships from scientific documents that can be used to populate structured chemical databases. Our system provides an extensible, chemistry-aware, natural language processing pipeline for tokenization, part-of-speech tagging, named entity recognition, and phrase parsing. Within this scope, we report improved performance for chemical named entity recognition through the use of unsupervised word clustering based on a massive corpus of chemistry articles. For phrase parsing and information extraction, we present the novel use of multiple rule-based grammars that are tailored for interpreting specific document domains such as textual paragraphs, captions, and tables. We also describe document-level processing to resolve data interdependencies and show that this is particularly necessary for the autogeneration of chemical databases since captions and tables commonly contain chemical identifiers and references that are defined elsewhere in the text. The performance of the toolkit to correctly extract various types of data was evaluated, arriving at an F-score of 93.4%, 86.8%, and 91.5% for extracting chemical identifiers, spectroscopic attributes, and chemical property attributes, respectively; set against the CHEMDNER chemical name extraction challenge, ChemDataExtractor yields a competitive F-score of 87.8%. All tools have been released under the MIT license and are available to download from <http://www.chemdataextractor.org>.



INTRODUCTION

Scientific results are typically communicated in the form of papers, patents, and theses that contain unstructured and semi-structured data described by free-flowing natural language that is not readily interpretable by machines. Yet, manual data abstraction by humans with expert knowledge is an expensive, labor-intensive, and error-prone process. With the continued growth of new publications, it is becoming increasingly difficult to create and maintain up-to-date manually curated databases, and automated information extraction by machines is fast becoming a necessity.

The chemistry literature presents an attractive and tractable target for this automated extraction as it is typically comprised of formulaic, data-rich language that is well-suited for machine analysis with the potential for high recall and precision. The extracted chemical information can be used to create and populate databases of chemical structures, properties, and observations, opening up new avenues for discovery through large-scale data mining studies that are of great value in diverse areas such as materials discovery, drug discovery, and intellectual property protection.

The chemistry literature presents an attractive and tractable target for this automated extraction as it is typically comprised of formulaic, data-rich language that is well-suited for machine analysis with the potential for high recall and precision. The extracted chemical information can be used to create and populate databases of chemical structures, properties, and observations, opening up new avenues for discovery through large-scale data mining studies that are of great value in diverse areas such as materials discovery, drug discovery, and intellectual property protection.

In recent years, efforts such as The Materials Genome Initiative¹ have led to an increased focus on large-scale data mining for materials discovery. Notable projects include the Harvard Clean Energy Project,² which focuses on materials for organic photovoltaics, and the Materials Project,³ which focuses

on battery materials. These existing projects are primarily confined to exploiting computational resources to predict chemical properties, an approach that will be well complemented by wider availability of machine-readable databases of experimental properties. Moreover, a generic method that can automatically generate a database for any type of material property would extend the reach of existing efforts to all areas of materials science, rather than predefining a focus on a specific area.

While there are many well-established text-mining tools in the biomedical domain,^{4–7} chemistry and materials text-mining is less widespread and fewer tools have been developed. Reviews by Elieb and Salim,⁸ Vazquez et al.,⁹ and Gunalingappa et al.¹⁰ provide comprehensive overviews of the existing text-mining tools and methods. Most of these tools focus narrowly on extracting specific entity types from specific document domains, while there are relatively few methodologies that embrace a broader focus on the extraction of chemical information, including properties, experimental measurements, and relationships between entities.

One such tool is ChemicalTagger,¹¹ which parses experimental synthesis sections of documents to determine chemical roles (e.g., reactant, solvent) and relationships with exper-

Received: April 13, 2016

Published: September 26, 2016

DOI: 10.1021/ci500207

J. Chem. Inf. Model. 2016, 56, 1894–1904

ACS Publications © 2016 American Chemical Society 1894

To tackle increasingly complex descriptions, scientists began using NLP tools

RESEARCH

ORGANIC CHEMISTRY

A universal system for digitization and automatic execution of the chemical synthesis literature

S. Hessam M. Mehr*, Matthew Craven*, Artem I. Leonov*, Graham Keenan, Leroy Cronin†

Robotic systems for chemical synthesis are growing in popularity but can be difficult to run and maintain because of the lack of a standard operating system or capacity for direct access to the literature through natural language processing. Here we show an extensible chemical execution architecture that can be populated by automatically reading the literature, leading to a universal autonomous workflow. The robotic synthesis code can be corrected in natural language without any programming knowledge and, because of the standard, is hardware independent. This chemical code can then be combined with a graph describing the hardware modules and compiled into platform-specific, low-level robotic instructions for execution. We showcase automated syntheses of 12 compounds from the literature, including the analgesic lidocaine, the Dess-Martin periodinane oxidation reagent, and the fluorinating agent AlkyFluor.

Synthetic chemistry remains labor intensive, and some protocols suffer from errors or ambiguity (1,2). Recently, there has been rapid growth in the development of robotic synthesis of molecules (3–5), but new developments are limited to specific reaction types, and a universal approach for the automatic encoding and validation of the chemical synthesis literature is lacking, which means that automation currently just displaces effort from manual labor to programming (6). The burden is further increased by the plethora of robotic solutions, which lack a common standard architecture. What is needed is an abstraction that can not only implement the literature (7) but also adapt to new synthetic methods (8–10), in accordance with a standard that ensures interoperability between hardware systems. Currently, the standard of the recording and subsequent reporting of the synthesis of new chemical compounds varies greatly, and procedures are often available only as incomplete and ambiguous prose, relying on the expert to fill in any gaps (12). This means that the quality of the data stored in many reaction databases is highly variable (13), posing many problems for reproducibility (14), as well as preventing the development of reliable digital methods for prediction of reactivity (15), new structures (16), and functions (17). These limitations have also prevented the practical digitization of chemistry (6)—i.e., the development of automated systems that could run reactions and make molecules—because of the lack of standards linking the reaction dependencies to a standard hardware control and specification, as well as a machine-readable standard for recording synthetic procedures.

Science Chemistry, University of Glasgow, Glasgow G12 8QQ, UK.

*These authors contributed equally to this work.

†Corresponding author. Email: lec.cronin@glasgow.ac.uk

Mehr *et al.*, *Science* **370**, 101–108 (2020) 2 October 2020

set of steps, with parallels between the manual actions in the traditional workflow and their automated counterparts in the digital workflow (Fig. 1B).

Comparing the two workflows reveals the steps necessary for realizing the digital workflow. These steps all operate on or refer to an intermediate representation of the actions in the procedure. To use a sufficiently capable batch synthesis robot for the universal digitization of chemistry, we thus hypothesized that the robot's hardware abstraction must be connected to practical synthesis by an executable, hardware-independent chemical programming language. This universal chemistry-oriented domain-specific markup language allows us to map the commonly reported methodological instructions to hardware-agnostic chemical instructions. Additionally, the semantic connection between this language and laboratory batch synthesis enables robotic execution of laboratory synthetic procedures without manual adaptation or modification. Instructions can thus be translated between the two without loss of information, as we no longer need to cross the large semantic gap between chemistry and robotic operations.

Herein, we present a system for the digitization of chemistry by making the chemical literature directly executable on automated synthesis platforms such as the Chemputer. To implement the system, we required a language with which synthetic procedures can be expressed succinctly by using vocabulary similar to that in the literature and executed on any compatible robotic platform. To this end, we have devised a chemical programming language and associated visual development environment that allows users to encode procedures without ambiguity and exchange them using a standard format. Our Chemical Description Language, XML (the Greek pronounced “XML” in reference to “χαρτό,” the Greek word for chemistry), achieves this goal by representing the synthesis procedures to these low-level instructions, which precludes portability to other platforms and requires programming expertise as well as detailed knowledge of the system's robotic operations. Such implementations demonstrate the capabilities of the hardware but are not a suitable or sustainable way of automating chemical synthesis. Far from being restricted to the Chemputer, the absence of a universal chemical programming language threatens to undermine the feasibility of the nascent automated synthesis ecosystem. Automated platforms from different companies or research groups all have bespoke instruction sets with no obvious semantic link among them or to the literature. This broken link has prevented the digitization of chemistry. Our vast repertoire of synthetic knowledge cannot be directly executed by robots today. To address this, we envisaged that a new architecture must rely on hardware-independent instructions represented in a standard chemical language that can express the synthesis of virtually any molecule (Fig. 1A).

The synthesis procedure, once written using our standard chemical programming language, provides a universal and hardware-independent way of digitizing chemical synthesis. However, there must also be a way of

The screenshot shows the article page for "Automated extraction of chemical synthesis actions from experimental procedures" published in Nature Communications. The page includes the journal logo, article type (ARTICLE), DOI (https://doi.org/10.1038/s41467-020-17266-6), and an "OPEN" access button. The main title is prominently displayed, followed by a brief abstract and author information. The text discusses the challenges of digitizing chemical synthesis and the development of a universal chemical programming language (Chemputer) to overcome these challenges. It highlights the use of a Chemical Description Language (XML) and machine learning to extract synthesis actions from experimental procedures. The article concludes with a summary of the Chemputer's capabilities and its potential impact on the field of synthetic chemistry.

¹IBM Research Europe, Säumerstrasse 4, Rüschlikon, 8803 Switzerland. ²These authors contributed equally. Alain C. Vaucher, Federico Zipoli. [✉]Email: ava@zurich.ibm.com

NATURE COMMUNICATIONS | (2020)11:3601 | https://doi.org/10.1038/s41467-020-17266-6 | www.nature.com/naturecommunications

1



Cite This: *Chem. Mater.* 2017, **29**, 2439–2444

Article
pubs.acs.org/cm

Materials Synthesis Insights from Scientific Literature via Text Extraction and Machine Learning

Edward Kim,^{1,2} Kevin Huang,¹ Adam Saunders,^{1,2} Andrew McCallum,^{1,2} Gerbrand Ceder,³ and Elsa Olivetti^{1,2,3,4,5}

¹Department of Materials Science and Engineering, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States

²Computer Science Department, University of Massachusetts Amherst, Amherst, Massachusetts 01003, United States

³Materials Science and Engineering, University of California Berkeley, Berkeley, California 94720, United States

⁴Supporting Information

ABSTRACT: In the past several years, Materials Genome Initiative (MGI) efforts have produced myriad examples of computationally designed materials in the fields of energy storage, catalysis, thermoelectrics, and hydrogen storage as well as large data resources that are used to screen for potentially transformative compounds. The bottleneck in high-throughput materials design has thus shifted to materials synthesis, which motivates our development of a methodology to automatically compile materials synthesis parameters across tens of thousands of scholarly publications using natural language processing techniques. To demonstrate our framework's capabilities, we examine the synthesis conditions for various metal oxides across more than 12 thousand manuscripts. We then apply machine learning methods to predict the critical parameters needed to synthesize titanias nanotubes via hydrothermal methods and verify this result against known mechanisms. Finally, we demonstrate the capacity for transfer learning by using our outperform heuristics strategies.

INTRODUCTION

First-principles materials design, open access materials property databases,^{1,2} and machine learning³ have accelerated novel compound identification for a variety of applications, including energy storage, catalysis, thermoelectrics, and hydrogen storage.^{4–10} To fully realize the vision of the Materials Genome Initiative of accelerating materials development,^{11–14} we must, in a comprehensive and accessible way, link the compositions, structures, and morphologies of these computationally discovered materials to the synthesis conditions that can produce them. This work represents a small step in the direction toward this goal of systematically understanding the relationships between synthesized materials and reaction conditions by broadly data mining the literature.

The materials design community remains gated by the use of heuristic synthesis guidelines; once a particular material of interest has been identified, either by direct first-principles computations or screening methods,^{6,15–20} as a result, the synthesis of targeted novel compounds is rapidly becoming the slow step in computationally driven materials design. With direct modeling of the complex kinetic processes occurring during synthesis out of reach, a data-driven, machine learning approach that learns from the hundreds of thousands of

published synthesis recipes may be more productive. As a step toward this objective, we use recent advances in full-text publisher application programming interfaces (APIs)²¹ and natural language processing (NLP)^{22–25} to develop a statistical learning approach to materials synthesis. While numerous studies have focused on text extraction from scientific literature,^{23–25,27–29} we present here a framework focused on the problem of extracting and data-mining materials synthesis conditions.

Using a variety of machine learning and natural language processing techniques, our platform automatically retrieves articles and their extracts and codifies the materials synthesis conditions and parameters found in the text. By combining these text-mined synthesis parameters at large scale, this synthesis database can be mined to discover the underlying relationships between synthesis conditions and the materials they produce. This literature-based data mining strategy also complements and benefits from current combinatorial and *in situ* synthesis studies which produce libraries of materials with

Received: August 21, 2017
Revised: October 8, 2017
Published: October 19, 2017



© 2017 American Chemical Society

9436

DOI: 10.1021/acs.chemmater.7b03500
Chem. Mater. 2017, **29**, 9436–9444



Large language models represented yet another jump forward in progress

SciBERT: A Pretrained Language Model for Scientific Text

Iz Beltagy Kyle Lo Arman Cohan
Allen Institute for Artificial Intelligence, Seattle, WA, USA
{beltagy,kylel,armanc}@allenai.org

Abstract

Obtaining large-scale annotated data for NLP tasks in the scientific domain is challenging and expensive. We release SciBERT, a pretrained language model based on BERT (Devlin et al., 2019) to address the lack of high-quality, large-scale labeled scientific data. SciBERT leverages unsupervised pretraining on a large multi-domain corpus of scientific publications to improve performance on downstream scientific NLP tasks. We evaluate on a suite of tasks including sequence tagging, sentence classification and dependency parsing, with datasets from a variety of scientific domains. We demonstrate statistically significant improvements over BERT and achieve new state-of-the-art results on several of these tasks. The code and pretrained models are available at <https://github.com/allenai/scibert>.

1 Introduction

The exponential increase in the volume of scientific publications in the past decades has made NLP an essential tool for large-scale knowledge extraction and machine reading of these documents. Recent progress in NLP has been driven by the adoption of deep neural models, but training such models often requires large amounts of labeled data. In general domains, large-scale training data is often possible to obtain through crowdsourcing, but in scientific domains, annotated data is difficult and expensive to collect due to the expertise required for quality annotation.

As shown through ELMO (Peters et al., 2018), GPT (Radford et al., 2018) and BERT (Devlin et al., 2019), unsupervised pretraining of language models on large corpora significantly improves performance on many NLP tasks. These models return contextualized embeddings for each token which can be passed into minimal task-specific

3615

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 3615–3620, Hong Kong, China, November 3–7, 2019. ©2019 Association for Computational Linguistics

JCIM JOURNAL OF CHEMICAL INFORMATION AND MODELING
pubs.acs.org/jcim

Automated Chemical Reaction Extraction from Scientific Literature
Jiang Guo,[§] A. Santiago Ibanez-Lopez,[§] Hanyu Gao, Victor Quach, Connor W. Coley, Klavs F. Jensen, and Regina Barzilay*

Cite This: *J. Chem. Inf. Model.* 2022, 62, 2035–2045 | Read Online

ACCESS | Metrics & More | Article Recommendations | Supporting Information

ABSTRACT: Access to structured chemical reaction data is of key importance for chemists in performing bench experiments and in modern applications like computer-aided drug design. Existing reaction databases are generally populated by human curators through manual abstraction from published literature (e.g., patents and journals), which is time consuming and labor intensive, especially with the exponential growth of chemical literature in recent years. In this study, we focus on developing automated methods for extracting reactions from chemical literature. We consider journal publications as the target source of information, which are more comprehensive and better represent the latest developments in chemistry compared to patents; however, they are less formulaic in their descriptions of reactions. To implement the reaction extraction system, we first devised a chemical reaction schema, primarily including a central *product*, and a set of associated reaction roles such as *reactants*, *catalyst*, *solvent*, and so on. We formulate the task as a structure prediction problem and solve it with a two-stage deep learning framework consisting of *product extraction* and *reaction role labeling*. Both models are built upon Transformer-based encoders, which are adaptively pretrained using domain and task-relevant unlabeled data. Our models are shown to be both effective and data efficient, achieving an F1 score of 76.2% in product extraction and 78.7% in role extraction, with only hundreds of annotated reactions.

INTRODUCTION
Scientific literature (e.g., journal articles and patents) has long been a critical information source to synthetic chemists for finding ways to perform particular chemical reactions or synthetic procedures of interest. To reduce the time and costs entailed by information retrieval, as well as to facilitate access to reaction data, commercial efforts have been invested in constructing structured databases from unstructured literature, such as Reaxys⁶ and SciFinder⁷ among others. These databases are generally populated by human experts through manual extraction from literature, which is costly, time consuming, and expertise intensive, especially with the exponential growth of scientific chemical publications in recent years.⁸ This challenge motivates the development of automated methods for reaction extraction from unstructured literature data.

Information extraction in the chemical domain has gained increasing attention over the past decade. Existing work has concentrated on named entity recognition (NER) and the extraction of their associated properties, such as OSCAR (Open-Source Chemistry Analysis Routines)⁸ and *ChemDataExtractor*.⁹ Only very few works have targeted the extraction of chemical reactions, which, compared to chemical compounds extracted by NER, are more structured, informative, and also practically useful. NER helps in associating compounds with documents, but chemists still need to go to the original article to see the context for that species, whereas reactions are often what the chemist wants to know about. Two representative toolkits developed for this

purpose are *ChemicalTagger*⁶ and OPSIN.⁷ *ChemicalTagger* went beyond entity extraction and used a grammar-based phrase parser to identify action phrases and relationships between entities. It has been specifically developed for extracting information from patents, taking advantage of its highly stylized and formulaic language. *OPSIN* took a mixture of outputs from *ChemicalTagger* and employed a set of rules to determine four essential chemical roles, including *product*, *reactant*, *solvent*, and *catalyst*. These rule-based systems represent good starting points for this endeavor, but they are heavily dependent on manually designed rules and are sensitive to the noise introduced by either language use or preprocessing steps, which limits their scalability to nonpatent data such as journal articles. Language used in academic journals is often of higher complexity and less formulaic than patent literature. For instance, one sentence can describe multiple reactions or one reaction with different products/yields under different conditions. This complexity requires the development of more advanced natural language processing (NLP) models with higher capacity. Another type of reaction data which is growing in popularity is synthesis action sequences, which

Special Issue: From Reaction Informatics to Chemical Space
Received: March 10, 2021
Published: June 11, 2021

2035

<https://doi.org/10.1021/acs.jcim.1c00284>
J. Chem. Inf. Model. 2022, 62, 2035–2045

npj computational materials

www.nature.com/npjcomputmat/

ARTICLE OPEN
MatSciBERT: A materials domain language model for text mining and information extraction

Tanishq Gupta,[§] Mohd Zaki,[§] N. M. Anoop Krishnan,^{§,2,3} and Mausam,^{§,4,5}

A large amount of materials science knowledge is generated and stored as text published in peer-reviewed scientific literature. While recent developments in natural language processing, such as Bidirectional Encoder Representations from Transformers (BERT) models, provide promising information extraction tools, these models may yield suboptimal results when applied on materials domain since they are not trained in materials science specific notations and jargons. Here, we present a materials-aware language model, namely, MatSciBERT, trained on a large corpus of peer-reviewed materials science publications. We show that MatSciBERT outperforms SciBERT, a language model trained on science corpus, and establish state-of-the-art results on three downstream tasks, named entity recognition, relation classification, and abstract classification. We make the pre-trained weights of MatSciBERT publicly accessible for accelerated materials discovery and information extraction from materials science texts.

INTRODUCTION

Discovering materials and utilizing them for practical applications is an extremely time-consuming process that may span decades.^{1,2} To accelerate this process, we need to exploit and harness the knowledge on materials that has been developed over the centuries through rigorous scientific procedure in a cohesive fashion.^{3–8} Textbooks, scientific publications, reports, handbooks, websites, etc., serve as a large data repository that can be mined for obtaining the already existing information.^{9,10} However, it is a challenging task to extract useful information from these texts since most of the scientific data is semi- or un-structured in the form of text, paragraphs with cross reference, image captions, and tables.^{11–12} Extracting such information manually is extremely time- and resource-intensive and relies on the interpretation of domain experts.

Natural language processing (NLP), a sub-domain in artificial intelligence, presents an alternative approach that can automate information extraction from text. Earlier approaches in NLP relied on non-neural methods based on n-grams such as Brown et al. (1992)¹³, structural learning framework by Ando and Zhang (2005)¹⁴, or structural correspondence learning by Blitzer et al. (2006)¹⁵, but these are no longer state of the art. Neural pre-trained embeddings like word2vec^{16,17} and GloVe¹⁸ are quite popular, but they lack domain-specific knowledge and do not produce contextual embeddings.

Recent progress in NLP has led to the development of a computational paradigm in which a large, pre-trained language model (LM) is finetuned for domain-specific tasks. Research has consistently shown that this pretrain-finetune paradigm leads to the best performance on downstream tasks.¹⁹ Statistically, LMs are probability distributions for a sequence of words such that for a given set of words, it assigns a probability to each word²⁰. Recently, due to the availability of large amounts of text and high computing power, researchers have been able to pre-train these large neural language models. For example, Bidirectional Encoder Representations from Transformers (BERT)²¹ is trained on BookCorpus²² and English Wikipedia, resulting in BERT²³ trained on 2 million clinical notes in MIMIC-III v1.4

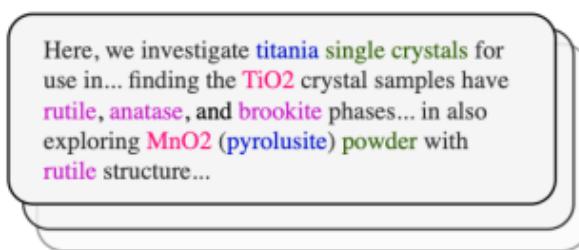
¹Department of Mathematics, Indian Institute of Technology Delhi, Hauz Khas, New Delhi 110016, India. ²Department of Civil Engineering, Indian Institute of Technology Delhi, Hauz Khas, New Delhi 110016, India. ³School of Artificial Intelligence, Indian Institute of Technology Delhi, Hauz Khas, New Delhi 110016, India. ⁴Department of Computer Science and Engineering, Indian Institute of Technology Delhi, Hauz Khas, New Delhi 110016, India. ⁵email: krishnan@iitd.ac.in; mausam@iitd.ac.in

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

npj

This grand challenge in data extraction seemed solvable with GPT

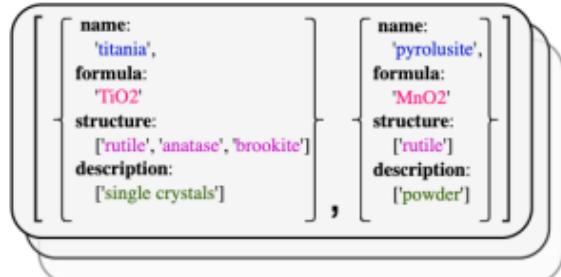
1. Training



Manual Annotation



~100 examples



2. Assisted Annotation

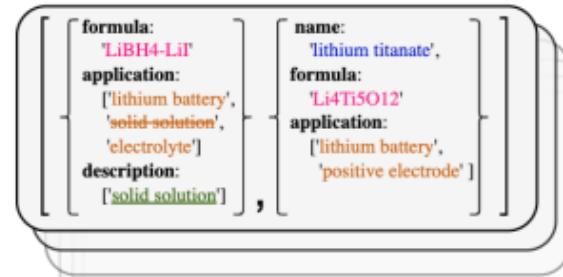
The charge and discharge performance of an all-solid-state lithium battery with the LiBH₄-LiI solid solution as an electrolyte is reported. Lithium titanate (Li₄Ti₅O₁₂) was used as the positive electrode and...

Training

Partially-Tuned LLM

Annotator Corrects Errors

~500 examples



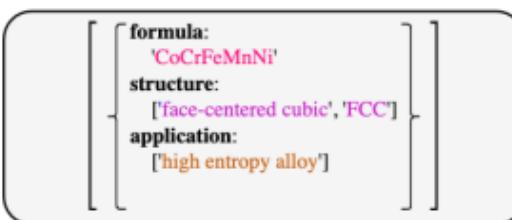
3. Inference

The equiatomic CoCrFeMnNi high entropy alloy, which crystallizes in the face-centered cubic (FCC) crystal structure, was prepared by the spark plasma sintering technique. Dynamic compressive tests of the ...

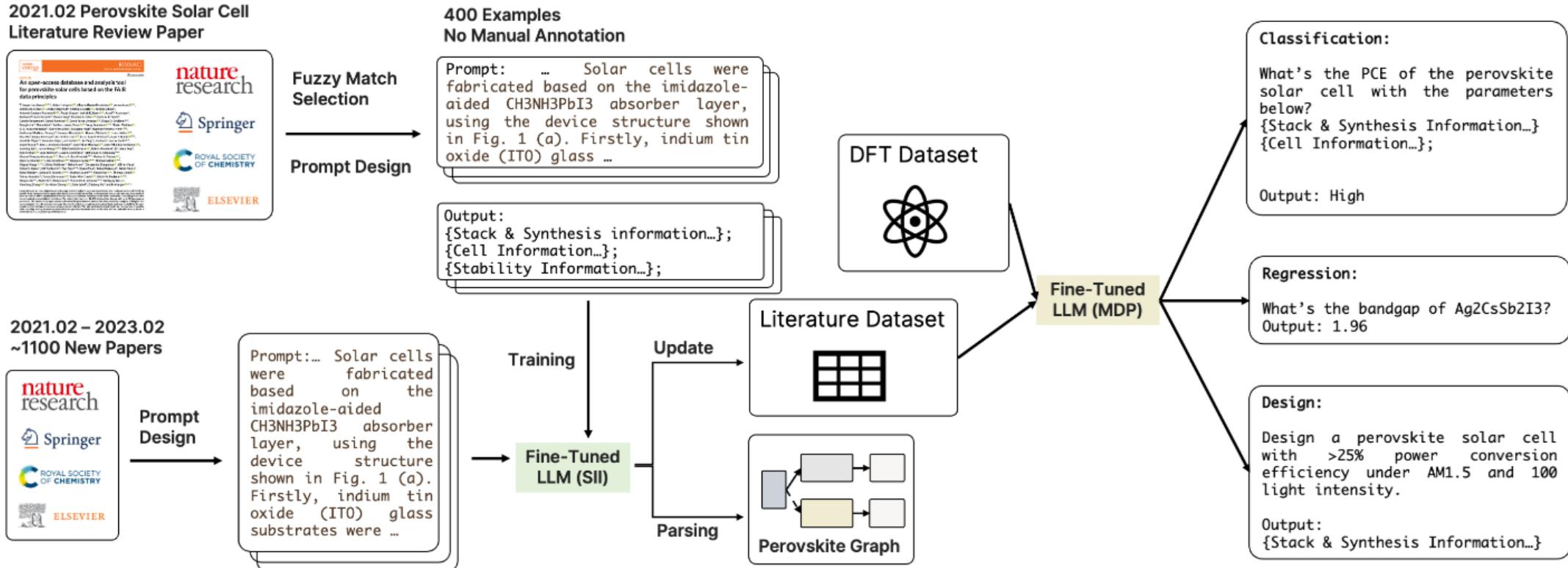
Training

Fine-Tuned LLM

Sequence Loss (Cross Entropy)



Now we see expansion of materials data extraction to specific areas



RLHF represented yet another drastic change in performance

Step 1

Collect demonstration data and train a supervised policy.

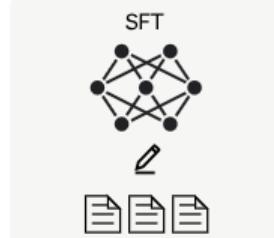
A prompt is sampled from our prompt dataset.



A labeler demonstrates the desired output behavior.



This data is used to fine-tune GPT-3.5 with supervised learning.



Step 2

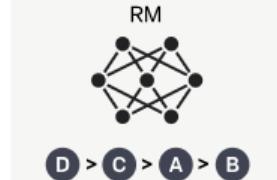
Collect comparison data and train a reward model.

A prompt and several model outputs are sampled.



A labeler ranks the outputs from best to worst.

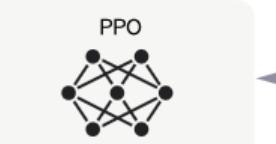
This data is used to train our reward model.



Step 3

Optimize a policy against the reward model using the PPO reinforcement learning algorithm.

A new prompt is sampled from the dataset.



The PPO model is initialized from the supervised policy.



The policy generates an output.

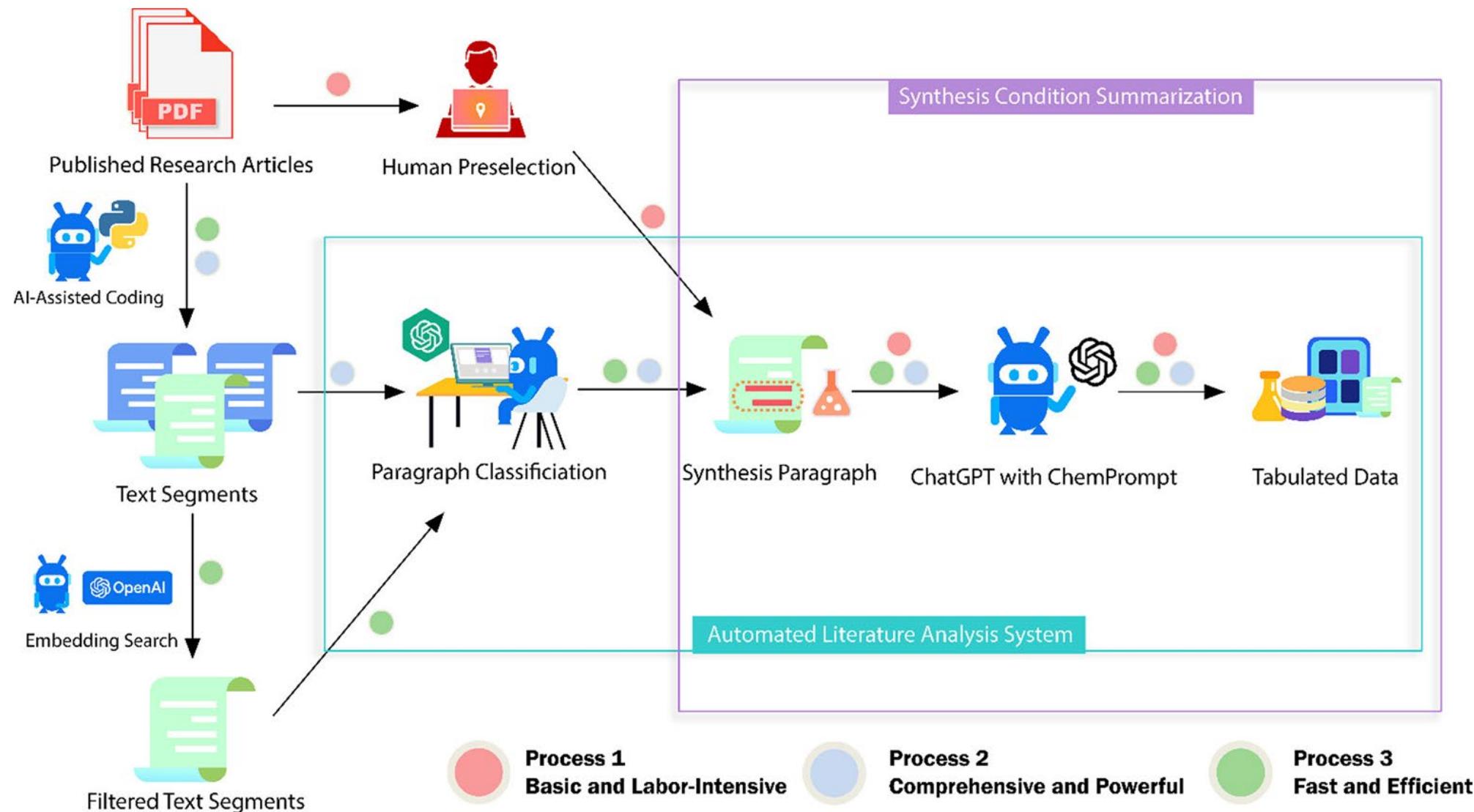


The reward model calculates a reward for the output.

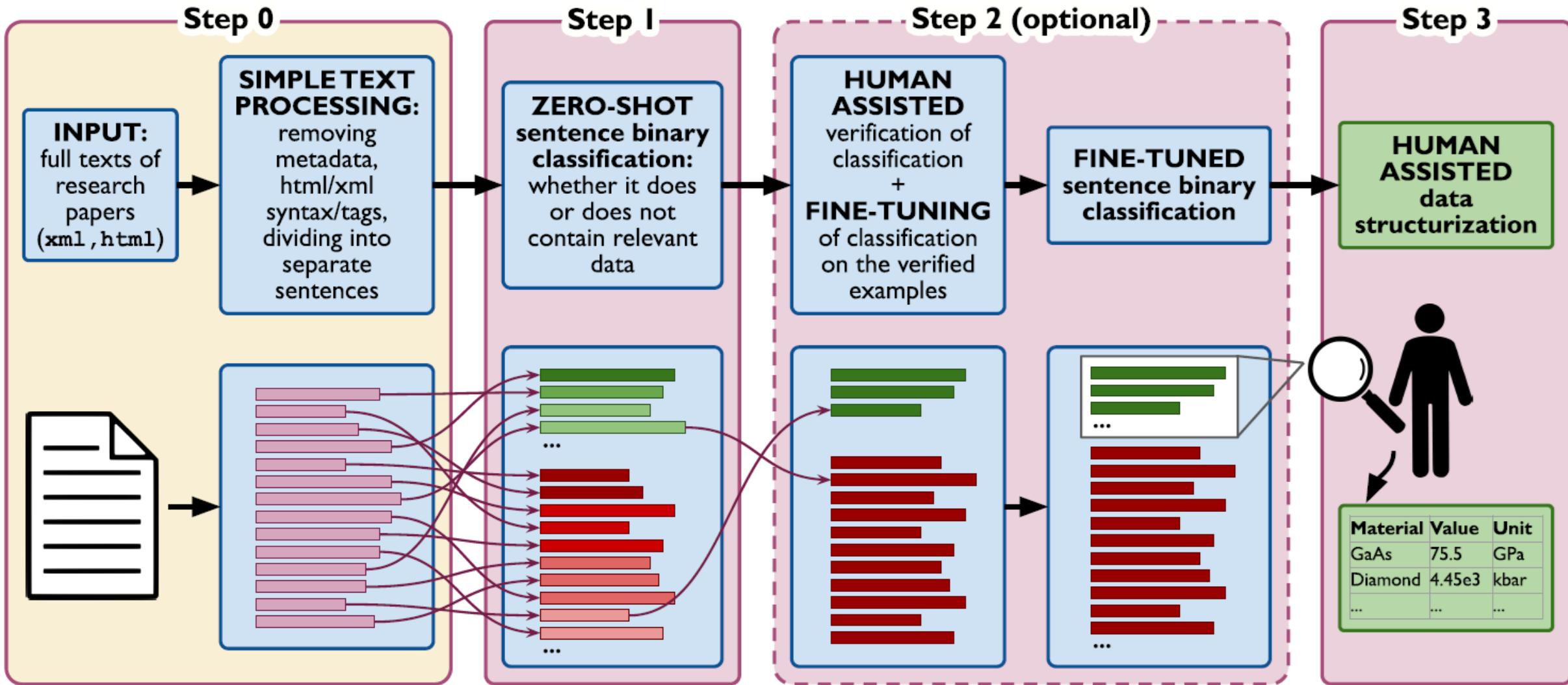


The reward is used to update the policy using PPO.

ChatGPT Chemistry Assistant for MOF Synthesis



Materials Data Extraction using general purpose LMs



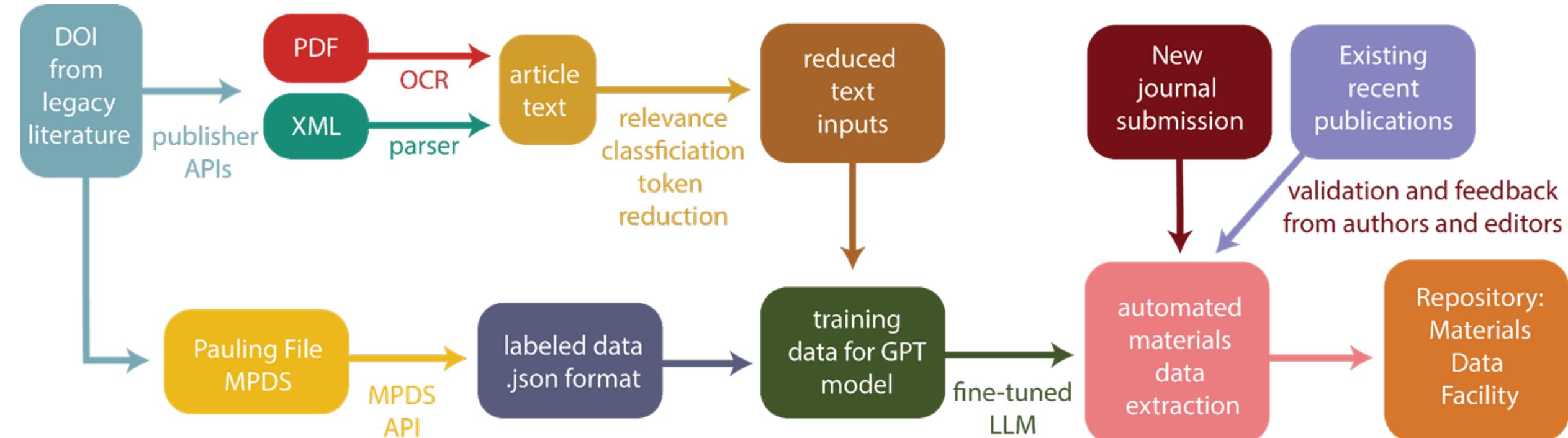
Jan 2023, we tested this on experimental sections with success

Substrate Preparation

A ~ 100 nm thick SiO₂ layer was grown on the surface of a p-type B-doped Si (100) wafer (Silicon Valley Microelectronics, Santa Clara, CA, USA) by plasma enhanced chemical vapour deposition (PECVD) with high radio frequency in a CVD MPX chamber (SPTS Technologies Ltd., Newport, UK). Afterwards, the wafer was cut in smaller portions (15 mm x 15 mm) whereon Ru-W films were co-deposited by DC magnetron sputtering in an ultra-high vacuum sputtering system (Kenosistec, Binasco, Italy). Power biases of 40-30 W were applied to Ru and W targets (99.95% purity, Testbourne Ltd., Basingstoke, UK), respectively. Deposition was performed in a chamber at 6.9 X 10⁻¹ Pa with Ar influx of 20 sccm for 600 s. Sputtering parameters were selected in order to produce a ~ 20 nm Ru-W film with near equimolar composition, based on previous work.

```
{  
  "substrate_preparation": {  
    "layer": "SiO2",  
    "thickness": "100 nm",  
    "substrate": "p-type B-doped Si (100) wafer",  
    "substrate_source": "Silicon Valley Microelectronics, Santa Clara, CA, USA",  
    "deposition_method": "plasma enhanced chemical vapour deposition (PECVD)",  
    "deposition_chamber": "CVD MPX chamber (SPTS Technologies Ltd., Newport, UK)",  
    "ruw_film_deposition_method": "DC magnetron sputtering",  
    "ruw_film_deposition_system": "ultra-high vacuum sputtering system (Kenosistec, Binasco, Italy)",  
    "ru_power_bias": "40 W",  
    "w_power_bias": "30 W",  
    "ru_purity": "99.95%",  
    "w_purity": "99.95%",  
    "ru_source": "Testbourne Ltd., Basingstoke, UK",  
    "w_source": "Testbourne Ltd., Basingstoke, UK",  
    "chamber_pressure": "6.9 × 10-1 Pa",  
    "ar_influx": "20 sccm",  
    "deposition_time": "600 s",  
    "ruw_film_thickness": "≈20 nm",  
    "ruw_film_composition": "near-equimolar"  
  },  
}
```

We see two key opportunities for improvement



Meta's Nougat works well for converting pdf into a markup language

2. Experimental

Co_2MnSn , $\text{Co}_2\text{MnGa}_{0.5}\text{Sn}_{0.5}$ and Co_2MnGa alloys were prepared by induction melting under an Ar atmosphere. To attain the $L2_1$ ordering, all the alloys were annealed in vacuum-sealed quartz tubes for 168 h at 673 K. This annealing temperature was selected to be lower than the $L2_1/\text{B}2$ order-disorder temperature estimated by differential thermal analysis (DTA) for Co_2MnGa [16]. The chemical compositions of the samples determined by induction coupled plasma (ICP) mass analysis are summarized in Table 1. The structures of the samples were investigated by the standard X-ray diffraction (XRD) technique using $\text{Cu } K_{\alpha}$ radiation, and transmission electron microscopy. (TEM) The spin-polarization measurements were carried out using the PCAR technique [17]. Sharp Nb tips were used to obtain the superconducting point contacts with the samples. As shown by Woods et al. and Perez-Willard et al., a ballistic model can be employed for diffusive contacts to estimate the spin polarization [18,19]. Therefore, the conductance curves were fitted with Strijkers model (ballistic) [20]. A multiple parameter fitting was carried out using the dimensionless interfacial scattering parameter Z , superconducting gap Δ and spin polarization P as the variable parameters. The *ab initio* DOS calculations were performed using the VASP 4.5 package, within generalized gradient approximations. In this new version, the k -points were set to auto for self-consistency, the cut-off energy of 290 eV was fixed, and the Perdew plane augmented pseudopotentials (GGA 91) were used to generate the total DOS [21,22]. Note that we have chosen the experimental lattice

2 Experimental

$\text{Co}(\{\}_{2})\text{MnSn}$, $\text{Co}(\{\}_{2})\text{MnGa}(\{\}_{0.5})\text{Sn}(\{\}_{0.5})$ and $\text{Co}(\{\}_{2})\text{MnGa}$ alloys were prepared by induction melting under an Ar atmosphere. To attain the $L2_1$ ordering, all the alloys were annealed in vacuum-sealed quartz tubes for 168 h at 673 K. This annealing temperature was selected to be lower than the $L2_1/\text{B}2$ order-disorder temperature estimated by differential thermal analysis (DTA) for $\text{Co}(\{\}_{2})\text{MnGa}$ [16]. The chemical compositions of the samples determined by induction coupled plasma (ICP) mass analysis are summarized in Table 1. The structures of the samples were investigated by the standard X-ray diffraction (XRD) technique using $\text{Cu } K_{\alpha}$ radiation, and transmission electron microscopy. (TEM) The spin-polarization measurements were carried out using the PCAR technique [17]. Sharp Nb tips were used to obtain the superconducting point contacts with the samples. As shown by Woods et al. and Perez-Willard et al., a ballistic model can be employed for diffusive contacts to estimate the spin polarization [18,19]. Therefore, the conductance curves were fitted with Strijkers model (ballistic) [20]. A multiple parameter fitting was carried out using the dimensionless interfacial scattering parameter Z , superconducting gap Δ and spin polarization P . The spin-polarization measurements were carried out using the PCAR technique [17]. Sharp Nb tips were used to obtain the superconducting point contacts with the samples. As shown by Woods et al. and Perez-Willard et al., a ballistic model can be employed for diffusive contacts to estimate the spin polarization [18,19]. Therefore, the conductance curves were fitted with Strijkers model (ballistic) [20]. A multiple parameter fitting was carried out using the dimensionless interfacial scattering parameter Z , superconducting gap Δ and spin polarization P .

Nougat: Neural Optical Understanding for Academic Documents

Paper arxiv.2308.13418 license MIT pypi v0.1.17 python 3.9+ code style black

 Hugging Face Community Space

This is the official repository for Nougat, the academic document PDF parser that understands LaTeX math and tables.

Project page: <https://facebookresearch.github.io/nougat/>

Nougat also works well on converting pdf tables

Table 1

$\text{Co}_2\text{MnSn}_{1-x}\text{Ga}_x$ alloy compositions, lattice parameters, ordering parameters S and α , and transition temperatures.

Alloy	Composition	Lattice parameter	S	α	T_c (K)	$T_{L21/B2}$ (K)	T_m (K)
Co_2MnSn	50.3:24.2:26.3	0.599	0.97	0.07	800	—	1370
$\text{Co}_2\text{MnGa}_{0.5}\text{Sn}_{0.5}$	50.2:24.0:12.3:12.5	0.589	0.92	0.03	770	—	1358
Co_2MnGa	50.6:24.3:25.1	0.577	0.98	0.45	700	1220	1440

```
\begin{table}
\begin{tabular}{ccccccccc} Alloy & Composition & Lattice parameter &  $S$  &  $\alpha$  &  $T_c$ (K) &  $T_{L21/n2}$ (K) &  $T_m$ (K) \\ \hline
\text{Co}\{2\}\text{MnSn} & 50.3:24.2-26.3 & 0.599 & 0.97 & 0.07 & 800 & — & 1370 \\
\text{Co}\{2\}\text{MnGa}\{0.5\}\text{Sn}\{0.5\} & 50.2:24.0:12.3:12.5 & 0.589 & 0.92 & 0.03 & 770 & — & 1358 \\
\text{Co}\{2\}\text{MnGa} & 50.6:24.3:25.1 & 0.577 & 0.98 & 0.45 & 700 & 1220 & 1440 \\ \end{tabular}
\end{table}

Table 1:  $\text{Co}\{2\}\text{MnSn}\{1-x\}\text{Ga}\{x\}$  alloy compositions, lattice parameters, ordering parameters  $S$  and  $\alpha$ , and transition temperatures.
```

Alloy	Composition	Lattice parameter	S	α	T_c (K)	$T_{L21/n2}$ (K)	T_m (K)
Co_2MnSn	50.3:24.2-26.3	0.599	0.97	0.07	800	—	1370
$\text{Co}_2\text{MnGa}_{0.5}\text{Sn}_{0.5}$	50.2:24.0:12.3:12.5	0.589	0.92	0.03	770	—	1358
Co_2MnGa	50.6:24.3:25.1	0.577	0.98	0.45	700	1220	1440

Table 1: $\text{Co}_2\text{MnSn}_{1-x}\text{Ga}_x$ alloy compositions, lattice parameters, ordering parameters S and α , and transition temperatures.

We did a preliminary study to find what type of data lives where in materials journals

Matter

Matter of Opinion

NLP meets materials science: Quantifying the presentation of materials data in literature

Hasan M. Sayeed,¹ Wade Smallwood,¹ Sterling G. Baird,^{1,2}
and Taylor D. Sparks^{1,*}

Large language models (LLMs) revolutionized how we engage with information. In materials science, we aim to leverage natural language processing to transform progress and discovery. Analyzing diverse materials science papers, we annotate data types and sources, laying the groundwork for targeted information extraction and LLM development.

Introduction

Materials science, an interdisciplinary field merging principles from physics, chemistry, and engineering, plays a pivotal role in technological advancement and societal progress. At its core, materials science focuses on the exploration and development of new materials. While the availability of structured data, particularly through advancements in machine learning, plays a pivotal role in expediting the discovery process,¹ materials science lags behind other physical sciences in this regard. This lag is mainly due to the diverse nature of data sources and the lack of centralization. The solution to this data challenge in materials science lies in making use of scientific publications, which are the primary means of communication in the field.² The extensive academic literature in this domain holds a wealth of data concerning material compositions, processing conditions, and performance properties. However, this valuable information is often scattered across the textual content, tables, and figures of research papers, posing a formidable challenge for data extraction and subsequent analysis.

Methods for data extraction in materials science encompass a combination of manual curation and automated tech-

niques. Manual curation offers precision but is labor intensive and struggles to keep pace with the growing volume of materials science literature. Conversely, automated techniques are still evolving, grappling with the intricacies and diversity of data found in materials science papers.³

This informal study aims to delve into the patterns of data expression in materials science papers and uncover the relationships between data sources, laying the foundation for more efficient data extraction methods as a stepping stone toward our broader objective: to extract data from materials science papers and train a natural language processing (NLP) model to transform these data into a more machine-readable format. The insights gained from this analysis will play a pivotal role in informing the development of our model and may contribute to more efficient data extraction and analysis in materials science research.

Method

We randomly selected ten research papers spanning diverse subfields within materials science. This selection ensured a representative sample, encompassing a wide range of topics from superionic conductors to lattice

anisotropy. The goal was to capture the breadth of data presentation practices in materials science literature.

The source or vehicle by which the data were presented within the papers was categorized by one of the following three definitions:

Text: This source refers to information presented within the narrative text of the research papers. It comprises descriptions, explanations, and discussions related to the materials, their compositions, processing conditions, and performance properties.

Tables: Organized tables in the papers were another key source of data. These tables typically offer structured and tabulated data, making it easier to access and interpret key information such as numerical values, experimental results, and comparative data.

Figures: Figures, including graphs, charts, diagrams, and micrographs, represent another significant data source. These visual representations often convey complex information succinctly and can include details about material properties, experimental results, and graphical depictions of scientific findings.

Within each paper, we focused on identifying four key data types or categories:

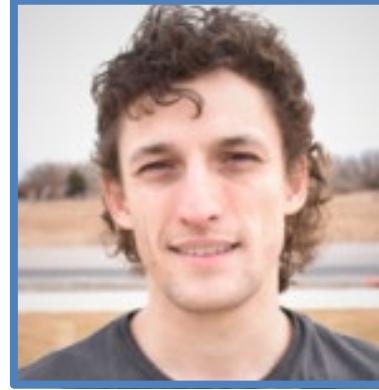
Chemical composition: Information related to the materials used, including chemical compositions, elements, and compounds.

¹Department of Materials Science and Engineering, University of Utah, 122 Central Campus Dr, Salt Lake City, UT 84112, USA

²Acceleration Consortium, University of Toronto, 80 St George St, Toronto, ON M5S 3H6, Canada

³Correspondence: sparks@engr.utah.edu

<https://doi.org/10.1016/j.matt.2023.12.032>



Text data is high variety and high complexity

Data Types:



Named Entity



Processing Condition



Characterization



Performance Properties



Text

$\text{Co}_2\text{MnGa}_{0.5}\text{Sn}_{0.5}$ were annealed in vacuum-sealed tubes for 168 h at 673 K. The structures ... were investigated by the



standard X-ray diffraction (XRD) technique using Cu K_α radiation, and transmission electron microscopy (TEM). The



...

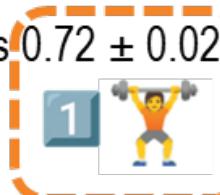
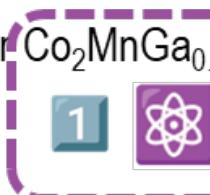
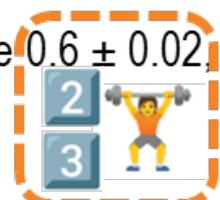


spin-polarization measurements were carried out using the PCAR technique. The spin-polarization values for Co_2MnGa

...



and Co_2MnSn were deduced to be 0.6 ± 0.02 , whereas that for $\text{Co}_2\text{MnGa}_{0.5}\text{Sn}_{0.5}$ was 0.72 ± 0.02 .



Tabular data is more structured, but less complete



Table

Data Types:



Named Entity



Processing Condition



Characterization



Performance Properties

	Alloy	Composition	Lattice Parameters	Ordering Parameter (S)	α	$T_c(K)$	$T_m(K)$
1	$Co_2MnGa_{0.5}Sn_{0.5}$	50.2:24.0:12.3:12.5	0.589	0.92	0.03	770	1358
2	Co_2MnGa	50.6:24.3:25.1	0.577	0.98	0.45	700	1440
3	Co_2MnSn	50.3:24.2:26.3	0.599	0.97	0.07	800	1370



Figure data is rich in data points, but lacking in categories

Data Types:



Named Entity



Processing Condition



Characterization



Performance Properties



Figure

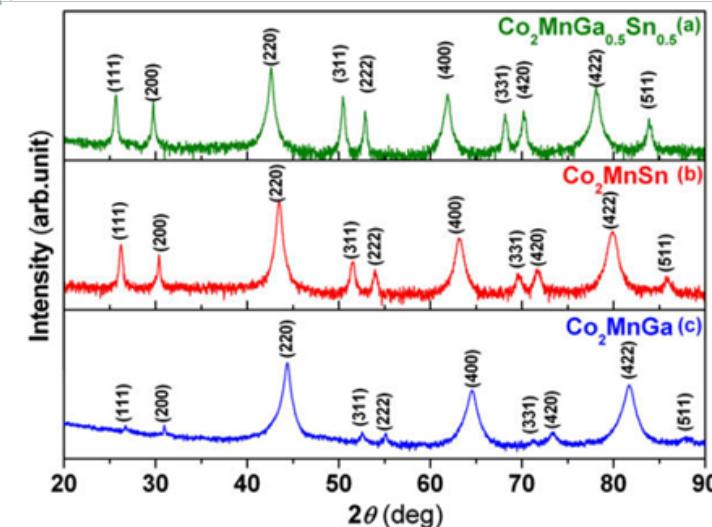


Fig. 3. XRD profiles of Co_2MnSn , $\text{Co}_2\text{MnGa}_{0.5}\text{Sn}_{0.5}$ and Co_2MnGa bulk alloys using $\text{Cu } K_{\alpha}$ radiation.



1

3

2

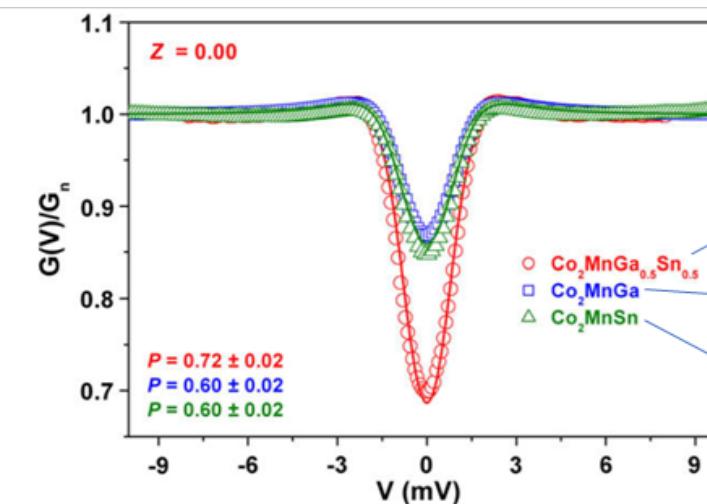


Fig. 7. The conductance curves of Co_2MnGa , $\text{Co}_2\text{MnGa}_{0.5}\text{Sn}_{0.5}$ and Co_2MnSn alloys, respectively, measured by PCAR at $T = 4.2$ K. Solid lines are the fits using the modified BTK model and open circles are the experimental data.

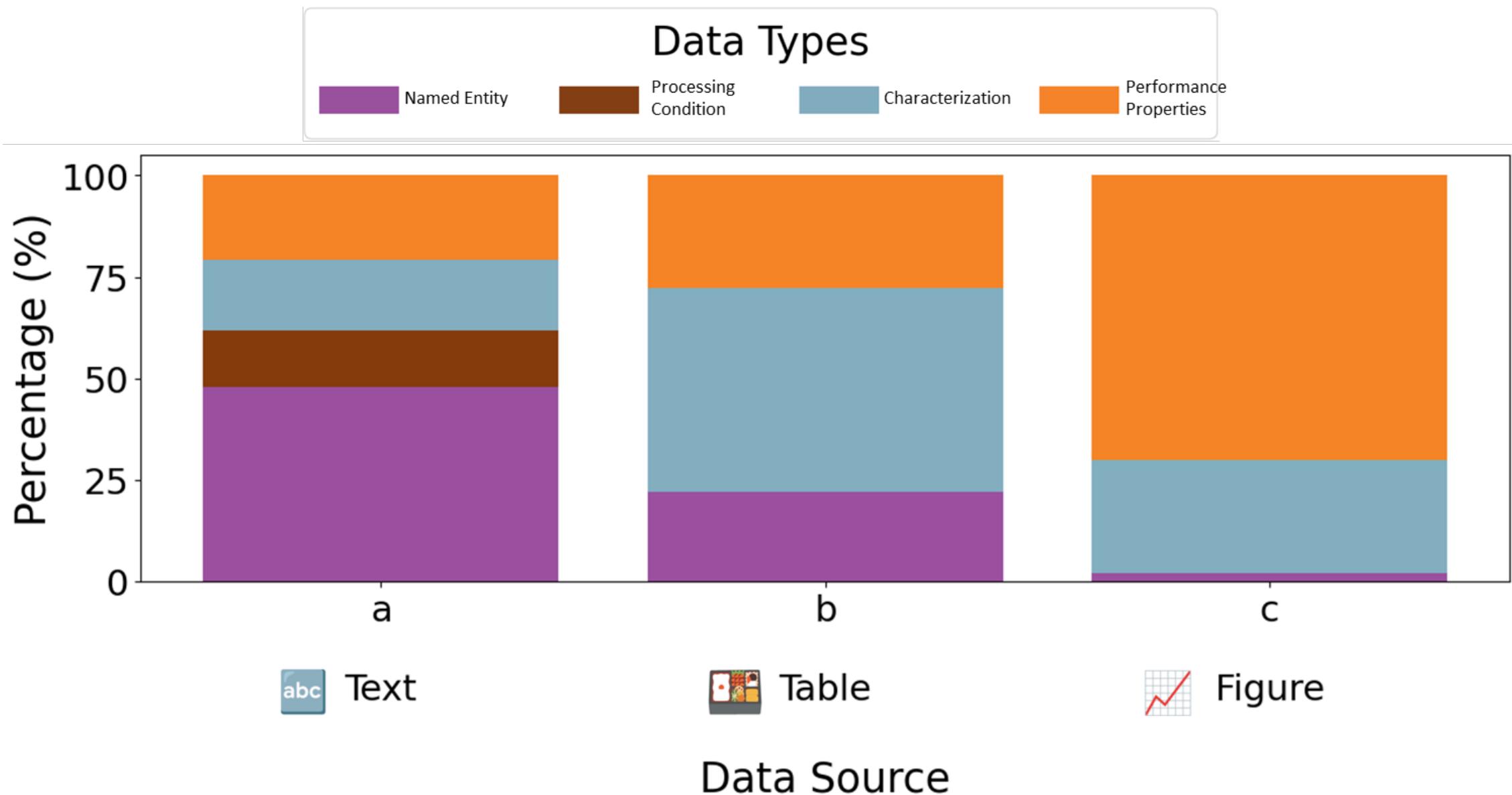


1

2

3

Different data is stored in text / table / figure. NLP ignores figures



Takeaway #1:
It's possible to go from PDF to text for mining but....
Not all the data is present in the text.

We have several options to choose from for token size reduction

Rules Based Extraction.

Ex. ChemDataExtractor

Extractive Summarization

Hand labeled data

0 Shot Summarization

No prompt engineering

1 Shot Summarization

Providing a single prompt



0 Shot Summarization Example

Summarize the following text:

Prompt

"Fig. 4a-c shows bright-field TEM images of the $\text{Co}_{(0.2)}\text{MnGa}$, $\text{Co}_{(0.2)}\text{MnGa}_{(0.5)}\text{Sn}_{(0.5)}$ and $\text{Co}_{(0.2)}\text{MnSn}$ alloys, respectively. The selected area diffraction patterns in the

Excerpt of Results Section of a Paper

Fig. 4a–c shows bright-field TEM images of the Co_2MnGa , $\text{Co}_2\text{MnGa}_{0.5}\text{Sn}_{0.5}$ and Co_2MnSn alloys, respectively. The selected area diffraction patterns in the insets confirm the presence of a L_2 phase, which is consistent with the XRD results. Antiphase boundaries (APBs) in the dark-field images excited with the (111) superlattice spots could not be observed in these three samples unlike other systems such as Co_2VAl [9]. The lack of APBs in these alloys indicates that the kinetics of the L_2 ordering during cooling from the melts and subsequent annealing was very fast, and therefore the L_2 ordered domains are much larger than the view of the TEM micrographs which indicates that the L_2 phase is stable in these alloys.

The magnetization curves of Co_2MnSn , $\text{Co}_2\text{MnGa}_{0.5}\text{Sn}_{0.5}$ and Co_2MnGa measured by SQUID at 10 K are shown in Fig. 5a. All of them show soft magnetic properties. The value of the saturation magnetization of Co_2MnGa , $\text{Co}_2\text{MnGa}_{0.5}\text{Sn}_{0.5}$ and Co_2MnSn is 4.04, 4.58 and $4.96 \mu_B/\text{f.u.}$, respectively. The magnetization values increase in accordance with the Curie temperatures (see

157 Tokens

58 Tokens



The text discusses the characterization and magnetic properties of three alloy systems: Co_2MnGa , $\text{Co}_2\text{MnGa}_{0.5}\text{Sn}_{0.5}$, and Co_2MnSn . Bright-field TEM images and selected area diffraction patterns confirm the presence of an L_2 phase in these alloys, suggesting rapid ordering during cooling and annealing. Antiphase boundaries are not observed in these samples, indicating the stability of the L_2 phase. The magnetization curves at 10 K show soft magnetic properties, and the saturation magnetization values increase with the Curie temperatures.



0 Shot Summarization Example with better instructions

Prompt

Make sure to concentrate on and provide any pertinent performance metrics, measurement results, or related numerical data while keeping the information concise.

Fig. 4a–c shows bright-field TEM images of the Co_2MnGa , $\text{Co}_2\text{MnGa}_{0.5}\text{Sn}_{0.5}$ and Co_2MnSn alloys, respectively. The selected area diffraction patterns in the insets confirm the presence of a L_2_1 phase, which is consistent with the XRD results. Antiphase boundaries (APBs) in the dark-field images excited with the (1 1 1) superlattice spots could not be observed in these three samples unlike other systems such as Co_2VAl [9]. The lack of APBs in these alloys indicates that the kinetics of the L_2_1 ordering during cooling from the melts and subsequent annealing was very fast, and therefore the L_2_1 ordered domains are much larger than the view of the TEM micrographs which indicates that the L_2_1 phase is stable in these alloys.

The magnetization curves of Co_2MnSn , $\text{Co}_2\text{MnGa}_{0.5}\text{Sn}_{0.5}$ and Co_2MnGa measured by SQUID at 10 K are shown in Fig. 5a. All of them show soft magnetic properties. The value of the saturation magnetization of Co_2MnGa , $\text{Co}_2\text{MnGa}_{0.5}\text{Sn}_{0.5}$ and Co_2MnSn is 4.04, 4.58 and $4.96 \mu_B/\text{f.u.}$, respectively. The magnetization values increase in accordance with the Curie temperatures (see

157
Tokens

50
Tokens



The text presents TEM images of Co_2MnGa , $\text{Co}_2\text{MnGa}_{0.5}\text{Sn}_{0.5}$, and Co_2MnSn alloys, confirming the presence of an L_2_1 phase. No antiphase boundaries are observed, indicating fast ordering kinetics and phase stability. Magnetization values at 10 K for Co_2MnGa , $\text{Co}_2\text{MnGa}_{0.5}\text{Sn}_{0.5}$, and Co_2MnSn are 4.04, 4.58, and $4.96 \mu_B/\text{f.u.}$, respectively, with increasing values corresponding to higher Curie temperatures.



0 Shot Approach varies 50% between good Metadata Extraction and Hallucination

Prompt

Excerpt of Experimental Section of a Paper

Please generate synthesis relevant metadata in JSON format from the following input:

All reagents and solvents used in this work were research grade products obtained from commercial sources (Merck-Sigma-Aldrich S.r.l., Milan, Italy) and used as received. The



```
Q {  
  "Reagents and Solvents": {  
    "source": "Merck-Sigma-Aldrich S.r.l., Milan, Italy",  
    "dinuclear copper complex": {  
      "name": "(HtomMe)(Cu(OAc))2",  
      "formula": "(C44H47Cu2N6O6)",  
      "description": "bis(acetate)-(μ-2,7-bis({bis[(6-methylpyridin-2-yl)methyl]amino}methyl)naphthalene-1,8-diolato)-di-copper(II)",  
      "synthesis": "synthesized, purified, and characterized, as described in the literature (19)"  
    },  
    "orthophosphoric acid": {  
      "name": "H3PO4",  
      "formula": null,  
      "description": null,  
      "range": "millimolar"  
    },  
    "1,2-ethylenediphosphonic acid": {  
      "name": "C2H8O6P2",  
      "formula": null  
    }  
  }  
}
```

```
Q {  
  "title": "Mass spectrometric and IRMPD spectroscopic studies of dinuclear copper complexes",  
  "authors": [  
    {  
      "name": "Lastname1, Firstname1",  
      "affiliation": "Department, University, City, Country"  
    },  
    {  
      "name": "Lastname2, Firstname2",  
      "affiliation": "Department, University, City, Country"  
    }  
  "journal": {  
    "name": "Journal Name",  
    "volume": "XX",  
    "issue": "X",  
    "pages": "XXX-XXX",  
    "year": "20XX"  
  }  
}
```

0 Shot approach is able to identify key categories intelligently

```
"methods": {  
    "reagents_and_solvents_source": "Merck-Sigma-Aldrich S.r.l., Milan, Italy",  
    "complex_synthesis_and_characterization_reference": "19",  
    "stock_acetonitrile_solution_composition": "C44H47Cu2N6O6",  
    "methanol_solutions_composition": [  
        "H3PO4",  
        "C2H8O6P2 (1,2-POH2)",  
        "C4H12O6P2 (1,4-POH2)",  
        "C10H14N5O6P (dAMP)",  
        "C10H14N5O7P (dGMP)"  
    ],  
    "reaction_mixture_preparation": {  
        "molar_ratio": "1:3",  
        "final_concentration": "(1–4) × 10–5 M for [1(OAc)2]+",  
        "dilution_solvent": "acetonitrile"  
    },  
    "analysis_method": "electrospray ionization (ESI) by direct infusion",  
    "analysis_details": {  
        "syringe_pump_flow_rate": "120 µL h-1"  
    }  
}
```

Chemical
Compounds

Processing
Conditions

Characterization

Hallucination: 0 Shot prompt: No Data extracted, rather a non-existent paper was organized

```
{  
  "title": "Mass spectrometric and IRMPD spectroscopic studies of dinuclear copper  
  complexes",  
  "authors": [  
    {  
      "name": "Lastname1, Firstname1",  
      "affiliation": "Department, University, City, Country"  
    },  
    {  
      "name": "Lastname2, Firstname2",  
      "affiliation": "Department, University, City, Country"  
    }  
  "journal": {  
    "name": "Journal Name",  
    "volume": "XX",  
    "issue": "X",  
    "pages": "XXX-XXX",  
    "year": "20XX"  
  }  
}
```

Takeaway #2

We need fine-tuned model for materials science data extraction. Generalized LLMs are not good enough for this.

To fine-tune we need gold-standard annotated dataset. That's time-consuming and expensive to do.

Pauling files from MPDS contains cherry picked data

Ni₃InSb

electrical resistivity	$\rho = 8.6 \cdot 10^{-7} \Omega \text{ m}$ [86 $\mu\Omega \text{ cm}$]	T= 330 K
thermoelectric power (Seebeck coefficient)	S= -8 $\mu\text{V K}^{-1}$	T= 330 K
thermoelectric figure of merit	$zT = 2.5 \cdot 10^{-3}$ [0.0025]	T= 330 K
power factor	$\sigma S^2 = 1.00 \cdot 10^{-4} \text{ W m}^{-1} \text{ K}^{-2}$ [100 $\mu\text{W m}^{-1} \text{ K}^{-2}$]	T= 330 K
thermal conductivity	$\lambda_{\text{th}} = 2.3 \text{ W m}^{-1} \text{ K}^{-1}$	T= rt
thermal conductivity	$\lambda_{\text{th}} = 11.5 \text{ W m}^{-1} \text{ K}^{-1}$	T= 330 K

Cancel

Send



Professional American football players



Cancel

Send



A group of people in American colonial



Cancel

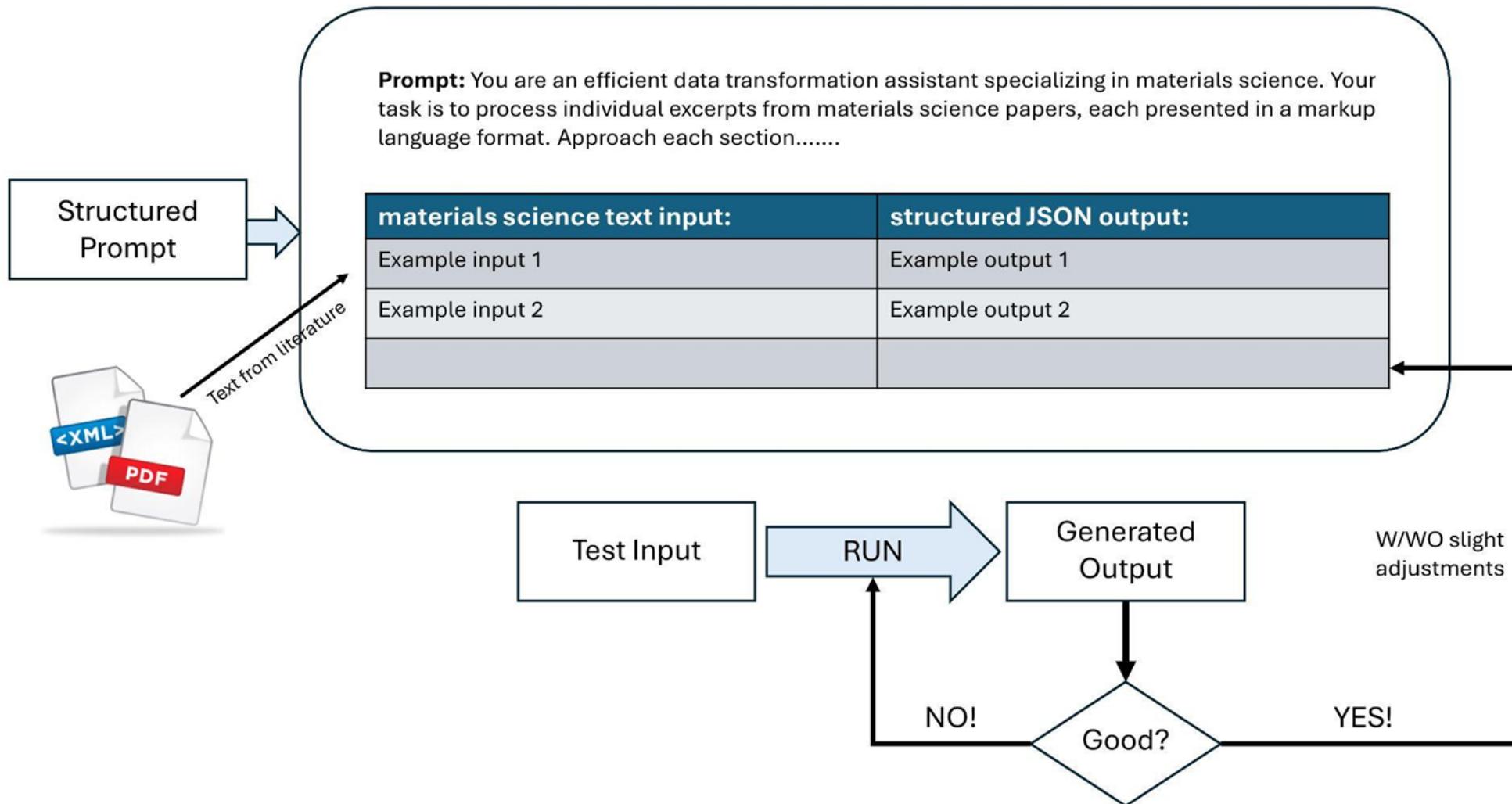
Send



Founding fathers



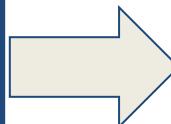
Google's Gemini Pro offers structured prompt workflow



Example of Abstractive Summary (Token Reduction)

Original Text:

In recent years, the field of materials science has made significant strides, particularly in the development of cathode materials for lithium-ion batteries. Among these, the synthesis of LiFePO₄ has garnered considerable attention due to its promising applications. The optimal synthesis process for LiFePO₄ involves heating the precursor materials at 800°C in an inert argon atmosphere for 10 hours. This high-temperature treatment is critical for ensuring the structural stability and purity of the final product. Interestingly, research in this area has often been compared to the culinary arts, where precise temperature and timing are crucial. The resulting LiFePO₄ material demonstrates enhanced electrochemical performance, characterized by a high specific capacity of 170 mAh/g and excellent cycle stability over 1000 cycles. The significance of these findings lies not just in their immediate application, but also in their potential to inspire further innovations in battery technology. The improved performance of LiFePO₄ can be attributed to its uniform particle size distribution and crystalline structure, which are essential for efficient lithium-ion transport. These advancements in cathode material synthesis could pave the way for more efficient, durable, and cost-effective energy storage solutions.

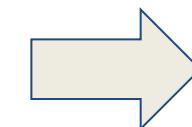


Extractive Summary:

"The optimal synthesis process for LiFePO₄ involves heating the precursor materials at 800°C in an inert argon atmosphere for 10 hours. The resulting LiFePO₄ material demonstrates enhanced electrochemical performance, characterized by a high specific capacity of 170 mAh/g and excellent cycle stability over 1000 cycles. The improved performance of LiFePO₄ can be attributed to its uniform particle size distribution and crystalline structure, which are essential for efficient lithium-ion transport."

Example of JSON Extraction

down to 20 K. It would be desirable to test these samples for superconductivity both resistively and magnetically at lower T , bearing in mind the 12 and 11 K superconductivity onsets in the related perovskite families $\text{Sr}_{1-x}\text{K}_x\text{BiO}_3-\delta$ [20] and $\text{Ba}_{1-x}\text{La}_x\text{PbO}_3-\delta$ [21]. The qualitative difference between the low- T thermal resistivity coefficients $d\rho/dT$ for $\text{Sr}_{1-x}\text{La}_x\text{PbO}_3-\delta$ and $\text{Sr}_{1-x}\text{K}_x\text{PbO}_3-\delta$ [17] may be due either to the bulk intrinsic asymmetry of electrons and holes in the parent substance $\text{SrPbO}_3-\delta$, or to the detrimental influence of the impurity phases on the conductivity in the inter-grain gap for $\text{Sr}_{1-x}\text{K}_x\text{PbO}_3-\delta$. Notwithstanding the metallic behavior of the polycrystalline $\text{Sr}_{1-x}\text{La}_x\text{PbO}_3-\delta$, the quantitative analysis of its resistivity must be performed taking into account the granular structure of the samples. The intrinsic $\rho(T)$ may be inferred only from single crystal investigations.



```
{  
  "chemical_composition": "Sr1-xKxBiO3-δ",  
  "processing_conditions": [],  
  "characterization": [],  
  "performance_properties": [  
    {  
      "property": "Superconductivity",  
      "value": "12",  
      "unit": "K"  
    }  
  ],  
  {  
    "chemical_composition": "Ba1-xLaxPbO3-δ",  
    "processing_conditions": [],  
    "characterization": [],  
    "performance_properties": [  
      {  
        "property": "Superconductivity",  
        "value": "11",  
        "unit": "K"  
      }  
    ]  
  },  
},  
]
```

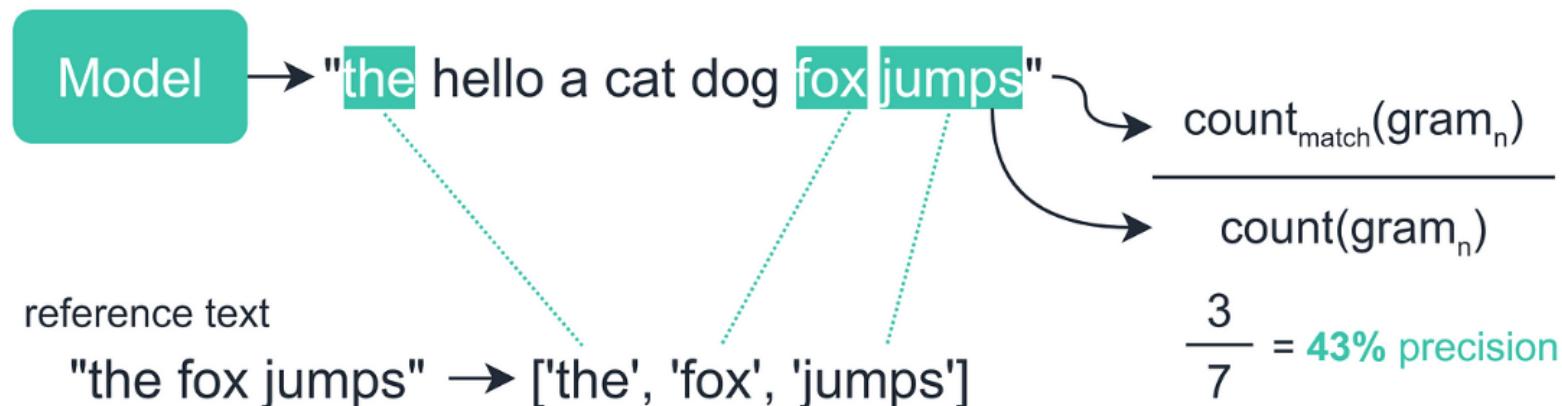
How good is structured prompt with Gemini Pro: testing

We tested performance on ten papers (entire text)

We ran structured prompts with only 5 examples going section by section and repeated the extraction 5 times

We compared the generated 5 outputs with a hand labeled reference to score performance

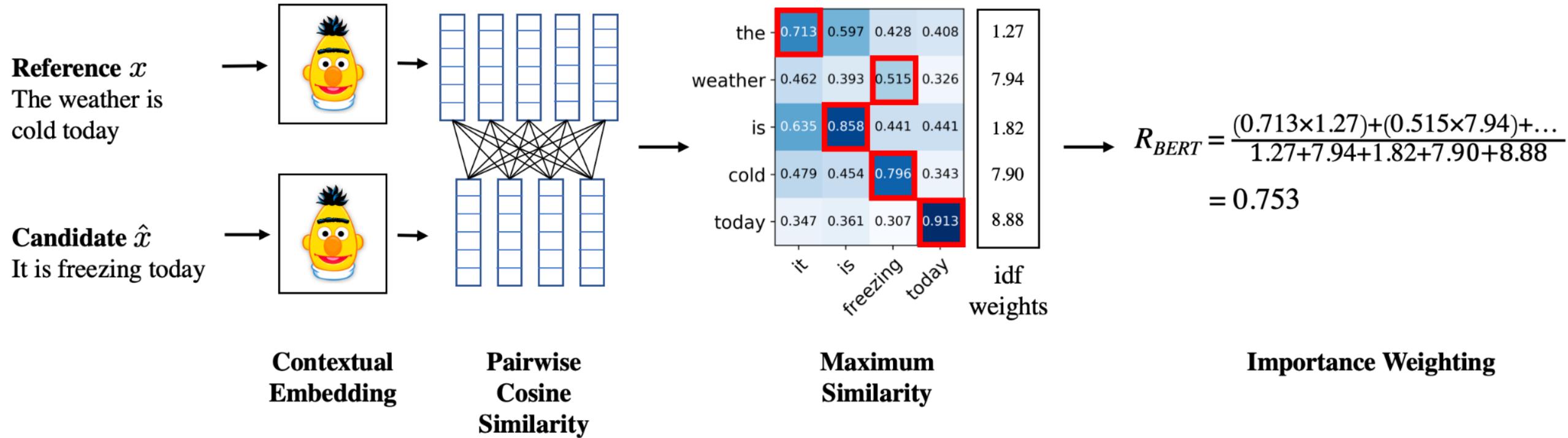
How good is structured prompt with Gemini Pro: ROUGEScore



$$2 * \frac{0.43 * 1.0}{0.43 + 1.0} = 0.6$$

60% f1 score

How good is structured prompt with Gemini Pro: BERTScore



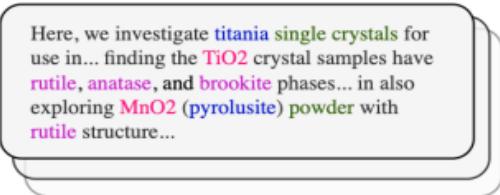
Structured prompts for data extraction is promising

Evaluation Metrics	Scores
ROUGE-1	0.58
ROUGE-2	0.36
ROUGE-L	0.56
BERTScore	0.92

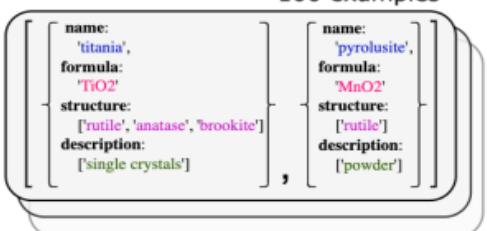
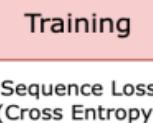
Takeaway #3

Structured prompts could greatly reduce the manual annotation step

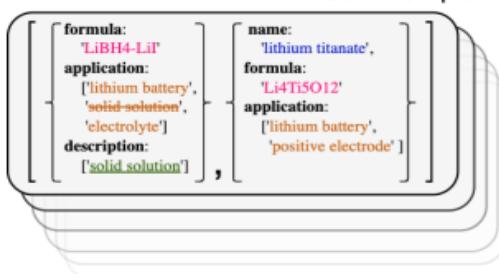
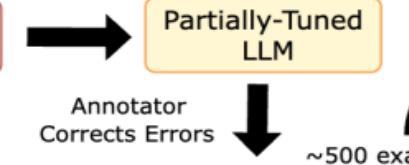
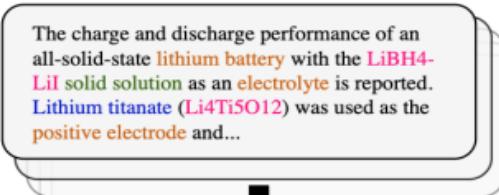
1. Training



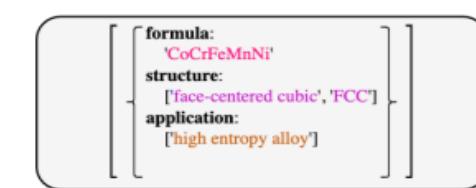
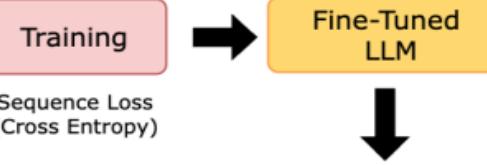
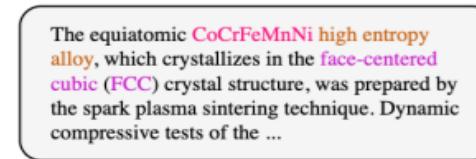
Manual Annotation



2. Assisted Annotation



3. Inference



Follow along with our developments!

Annotating Materials Science Text: A Semi-Automated Approach for Crafting Outputs with Gemini Pro

Hasan M. Sayeed¹, Trupti Mohanty¹, and Taylor D. Sparks *¹

¹Department of Materials Science & Engineering, University of Utah, Salt Lake City, UT 84112, USA

Abstract

Recent advancements in large language models (LLMs) have paved the way for automated information extraction in the materials science domain. However, fine-tuning these models, crucial for effective machine learning pipelines in materials science, is hindered by a lack of pre-annotated data. Manual annotation, a laborious process, exacerbates the challenge. To address this, we introduce a tailored semi-automated annotation process, using Google's Gemini Pro language model. Our approach focuses on two key tasks: extracting information in structured JSON format and generating abstractive summaries from materials science texts. The collaborative process, a symbiotic effort between human annotators and the LLM, driven by structured prompts and user-guided examples, enhances the annotation quality and augments the LLM's capacity to comprehend materials science intricacies. Importantly, it streamlines human annotation efforts by leveraging the LLM's proficient starting point.

1 Introduction

In the realm of materials science research, where a vast repository of knowledge is encapsulated within peer-reviewed scientific literature, the extraction of structured information and concise summaries plays a pivotal role in advancing the field. The intricate details and discoveries within these texts hold the key to unlocking new insights and driving innovation. Structured information, organized in a machine-readable format, not only facilitates efficient data retrieval but also lays the groundwork for systematic analysis and comparison. Similarly, concise summaries distill the essence of comprehensive research articles, offering researchers, scientists, and industry professionals a quick and insightful overview.

A significant portion of experimental data in materials science remains locked within the confines of literature [1, 2]. Unlocking this wealth of high-quality data presents an opportunity to revolutionize materials informatics, potentially leading to the creation of databases that could, overnight, transform the field. Indeed, such databases have already demonstrated their utility in constructing models capable of identifying regions in materials composition space conducive to superconductivity [3], designing high-entropy alloys [4], predicting the emergence of novel magnets [5], and forecasting the ZT thermoelectric figure of merit in inorganic materials [6], among other applications.

However, existing databases in the field often revolve around calculated materials properties [7–10], introducing potential systematic errors and limiting the scope of data to quantities amenable to rapid computations. Additionally, there exists the risk that such calculations may not accurately reflect real-world scenarios. This suboptimal reliance on calculated properties highlights the need for more comprehensive and accessible experimental databases. While some attempts at constructing experimental databases exist, they often remain proprietary due to the associated labor costs [11–13]. Although a few open-access initiatives have initiated database construction efforts [14, 15], there is still ample room for the development of workflows aimed at enhancing accessibility and performance in automated materials database construction from literature – a fertile area for investigation. Our JSON structured data extraction method emerges as a promising avenue in this exploration.

Previous attempts to extract data from literature showcase the potential for automation. Pipelines facilitating automated extraction of compound-property relationships from unstructured battery-specific texts [16] and the extraction of a database of 300,000 polymer property records from 650,000 abstracts [17] exemplify the strides made. Fine-tuned GPT models have been employed to extract general chemistry information from literature abstracts [18], demonstrating the versatility of such approaches. Domain-specific pre-training and fine-tuning have proven effective in enhancing transformer-based models' performance in domain-specific tasks [19–21]. As the emerging large language models (LLMs) are transformer-based, optimizing LLM pipelines for relevant domains becomes imperative, necessitating annotated data for fine-tuning.

Recent advancements in LLMs have presented exciting prospects for automating information extraction pipelines. Fine-tuning LLMs is crucial for constructing effective pipelines, especially in materials science. However, the scarcity of pre-annotated data poses a formidable challenge

*Corresponding author: sparks@engr.utah.edu

1



LLMs can find patterns in words, and those words can be properties!



Computer Science > Computation and Language

[Submitted on 21 Oct 2023]

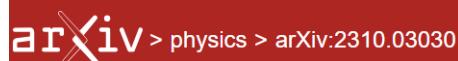
LLM-Prop: Predicting Physical And Electronic Properties Of Crystalline Solids From Their Text Descriptions

Andre Niyongabo Rubungo, Craig Arnold, Barry P. Rand, Adji Bouso Dieng

The prediction of crystal properties plays a crucial role in the crystal design process. Current methods for predicting crystal properties focus on modeling crystal structures using graph neural networks (GNNs). Although GNNs are powerful, accurately modeling the complex interactions between atoms and molecules within a crystal remains a challenge. Surprisingly, predicting crystal properties from crystal text descriptions is understudied, despite the rich information and expressiveness that text data offer. One of the main reasons is the lack of publicly available data for this task. In this paper, we develop and make public a benchmark dataset (called TextEdge) that contains text descriptions of crystal structures with their properties. We then propose LLM-Prop, a method that leverages the general-purpose learning capabilities of large language models (LLMs) to predict the physical and electronic properties of crystals from their text descriptions. LLM-Prop outperforms the current state-of-the-art GNN-based crystal property predictor by about 4% in predicting band gap, 3% in classifying whether the band gap is direct or indirect, and 66% in predicting unit cell volume. LLM-Prop also outperforms a finetuned MatBERT, a domain-specific pre-trained BERT model, despite having 3 times fewer parameters. Our empirical results may highlight the current inability of GNNs to capture information pertaining to space group symmetry and Wyckoff sites for accurate crystal property prediction.

Search...
Help

LLM
outperforms
GNN!?



Physics > Chemical Physics

[Submitted on 20 Sep 2023 (v1), last revised 10 Oct 2023 (this version, v3)]

GPT-MoLBERTa: GPT Molecular Features Language Model for molecular property prediction

Suryanarayanan Balaji, Rishikesh Magar, Yayati Jadhav, Amir Barati Farimani

With the emergence of Transformer architectures and their powerful understanding of textual data, a new horizon has opened up to predict the molecular properties based on text description. While SMILES are the most common form of representation, they are lacking robustness, rich information and canonicity, which limit their effectiveness in becoming generalizable representations. Here, we present GPT-MoLBERTa, a self-supervised large language model (LLM) which uses detailed textual descriptions of molecules to predict their properties. A text based description of 326000 molecules were collected using ChatGPT and used to train LLM to learn the representation of molecules. To predict the properties for the downstream tasks, both BERT and RoBERTa models were used in the finetuning stage. Experiments show that GPT-MoLBERTa performs well on various molecule property benchmarks, and approaching state of the art performance in regression tasks. Additionally, further analysis of the attention mechanisms show that GPT-MoLBERTa is able to pick up important information from the input textual data, displaying the interpretability of the model.

Search...
Help |

Better
representation
than SMILES?!

LLMs can find patterns in words, and those words can be properties!

nature machine intelligence



Article

<https://doi.org/10.1038/s42256-023-00788-1>

Leveraging large language models for predictive chemistry

Received: 16 May 2023

Kevin Maik Jablonka^{1,2,3,4}, Philippe Schwaller⁵, Andres Ortega-Guerrero⁶ & Berend Smit¹✉

Accepted: 22 December 2023

Published online: 6 February 2024

Check for updates

Machine learning has transformed many fields and has recently found applications in chemistry and materials science. The small datasets commonly found in chemistry sparked the development of sophisticated machine learning approaches that incorporate chemical knowledge for each application and, therefore, require specialized expertise to develop. Here we show that GPT-3, a large language model trained on vast amounts of text extracted from the Internet, can easily be adapted to solve various tasks in chemistry and materials science by fine-tuning it to answer chemical questions in natural language with the correct answer. We compared this approach with dedicated machine learning models for many applications spanning the properties of molecules and materials to the yield of chemical reactions. Surprisingly, our fine-tuned version of GPT-3 can perform comparably to or even outperform conventional machine learning techniques, in particular in the low-data limit. In addition, we can perform inverse design by simply inverting the questions. The ease of use and high performance, especially for small datasets, can impact the fundamental approach to using machine learning in the chemical and material sciences. In addition to a literature search, querying a pre-trained large language model might become a routine way to bootstrap a project by leveraging the collective knowledge encoded in these foundation models, or to provide a baseline for predictive tasks.

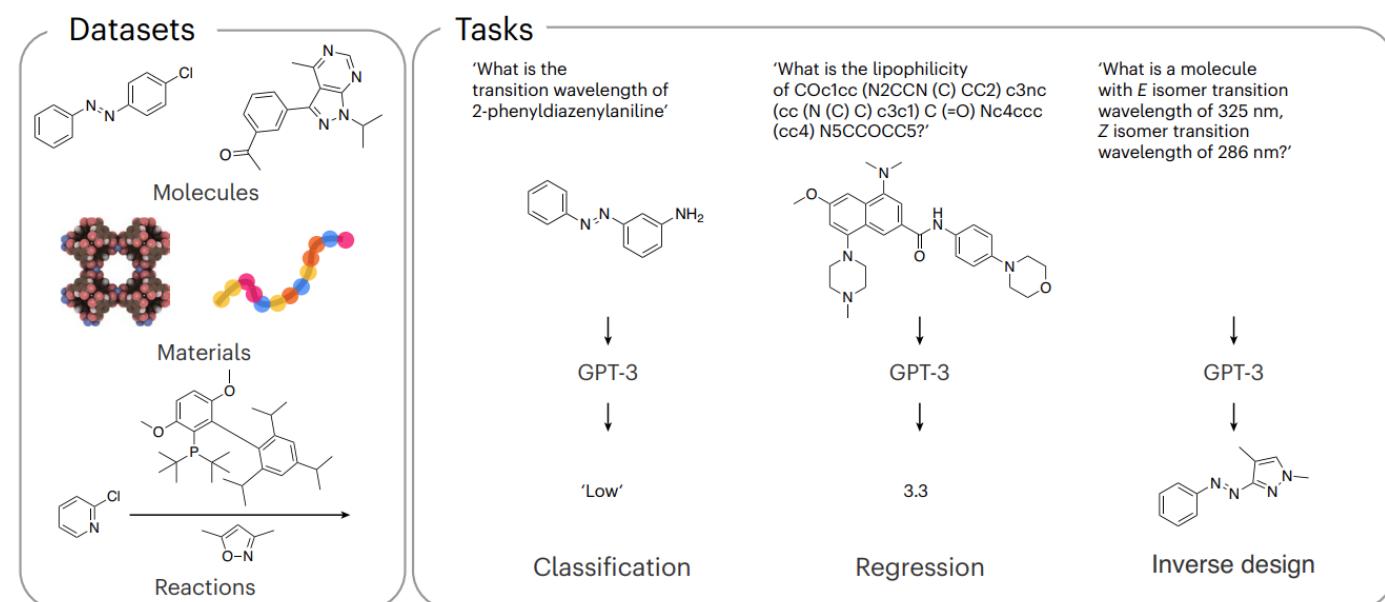
One of the fascinating advances in machine learning has been the development of large language models (LLMs), so-called foundation models^{1–4}. These models are appealing because of their simplicity; given a phrase, they return text that completes phrases in natural language such that, in many instances, one cannot tell that a machine wrote it.

From a scientific point of view, the most striking examples are that these foundation models can write sensible abstracts for scientific articles or even code for particular programming tasks^{5–11}. Recently, it has been shown that these models can also

solve relatively simple tabular regression and classification tasks¹³. However, as these models were not explicitly trained on these tasks, it is a remarkable result¹⁴.

That these models can solve simple tasks they are not trained for made us wonder whether they can also answer scientific questions for which we do not have an answer. As most chemistry problems can be represented in text form, we should be able to train these models to answer questions that chemists have. For example, 'If I change the metal in my metal–organic framework, will it be stable in water?'

¹Laboratory of Molecular Simulation (LMSO), Institut des Sciences et Ingénierie Chimiques, École Polytechnique Fédérale de Lausanne (EPFL), Sion, Switzerland. ²Center for Energy and Environmental Chemistry Jena (CEEC Jena), Friedrich Schiller University Jena, Jena, Germany. ³Laboratory of Organic and Macromolecular Chemistry (LOMC), Friedrich Schiller University Jena, Jena, Germany. ⁴Helmholtz Institute for Polymers in Energy Applications, Jena, Germany. ⁵Laboratory of Artificial Chemical Intelligence (LIAC), École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland. [✉]e-mail: berend.smit@epfl.ch



We can even use the word generation to create new structures!

Crystal Structure Generation with Autoregressive Large Language Modeling

Luis M. Antunes^{*1}, Keith T. Butler², and Ricardo Grau-Crespo^{*1}

¹Department of Chemistry, University of Reading, Whiteknights, Reading RG6 6DX, United Kingdom.
l.m.antunes@pgr.reading.ac.uk; r.grau-crespo@reading.ac.uk

²Department of Chemistry, University College London, WC1H 0AJ, United Kingdom.

Abstract

The generation of plausible crystal structures is often the first step in predicting the structure and properties of a material from its chemical composition. Quickly generating and predicting inorganic crystal structures is important for the discovery of new materials, which can target applications such as energy or electronic devices. However, most current methods for crystal structure prediction are computationally expensive, slowing the pace of innovation. Seeding structure prediction algorithms with quality generated candidates can overcome a major bottleneck. Here, we introduce CrystaLLM, a methodology for the versatile generation of crystal structures, based on the autoregressive large language modeling (LLM) of the Crystallographic Information File (CIF) format. Trained on millions of CIF files, CrystaLLM focuses on modeling crystal structures through text. CrystaLLM can produce plausible crystal structures for a wide range of inorganic compounds unseen in training, as demonstrated by *ab initio* simulations. The integration with predictors of formation energy permits the use of a Monte Carlo Tree Search algorithm to improve the generation of meaningful structures. Our approach challenges conventional representations of crystals, and demonstrates the potential of LLMs for learning effective ‘world models’ of crystal chemistry, which will lead to accelerated discovery and innovation in materials science.

1 Introduction

The *in silico* search for new materials often involves the exploration of a space of compositions in a chemical system, and the investigation of various predicted structural phases in that space (see [1], [2] and [3] for examples). To elucidate the structures of unknown materials, a Crystal Structure Prediction (CSP) approach is often employed, which attempts to derive the ground state crystal structure for a given chemical composition under specific physical conditions. [4] CSP approaches are relatively computationally expensive, typically involving *ab initio* techniques. [5] They often begin with the generation of candidate structures. Examples are the AIRSS [6, 7] and USPEX [8] approaches. Initializing the search space with sensible structures increases the likelihood of success, and decreases the

