International Journal of Advance and Innovative Research
Volume 12, Issue 3: July - September 2025
ISSN 2394 - 7780

EVALUATING AI IN MATERIAL SCIENCE

Sahil Sekhri
Director
Golden dots international Pvt Ltd

ABSTRACT
The rapid advancement of artificial intelligence has introduced significant potential for data-driven research in materials science. This study investigates the capabilities of Large Language Models (LLMs) in extracting and rea- soning with complex materials science information. We assess the performance of models such as GPT-3.5-Turbo and GPT-4 in named entity recognition (NER) and relation extraction (RE) tasks, focusing on their ability to handle domain-specific materials and property expressions. Using datasets like MeasEval and SuperMat, we compare these LLMs against traditional rule-based approaches and BERT-based models, establishing a baseline through precision, recall, and F1-score metrics. Our novel evaluation method, which normalizes materials to their chemical formulas for pairwise element comparison, reveals the strengths and limitations of LLMs in materials science. While LLMs demonstrate proficiency in general tasks, our results highlight the challenges they face in tasks requiring deep domain knowledge and the necessity for further refinement in this field. This research contributes to the understanding of how AI can be harnessed to accelerate the discovery and design of novel materials, marking a step towards more efficient and automated processes in materials science.

I. INTRODUCTION

Mining information from the scientific literature is gaining momentum in material science because of its availability and other attempts to increase its usage. Data for AI in materials science is often sourced from research papers, databases, lab experiments, or first-principles calculations [1]. The use of big data in this field has shifted from random meth- ods to more efficient, data-driven techniques. For instance, mining computational screening libraries has identified CO_2- binding sites, aiding in the discovery of materials with specific properties for wet gas environments [2]. Machine learning has been applied in high-entropy alloy discovery, leveraging probabilistic models and neural networks [3]. However, a key limitation in advanced AI for exploratory materials science is the lack of sufficiently large and diverse datasets suitable for data mining [4]. A central tenet in data-driven materials discovery is that sufficient data and appropriate techniques could significantly streamline new material development [5]. Materials science is moving from traditional manual processes to automated, parallel, and iterative processes driven by AI, simulation, and experimental automation [6], [7]. Despite the availability of large text-based information in materials science literature, it remains underutilized due to the challenges of extracting data from diverse formats such as unstructured text, tables, and figures [8]. Consequently, much data extraction still relies on manual effort. While structured databases exist, they are limited and expensive to maintain due to the labor-intensive curation required [9], [10].

Challenges: One significant challenge is ensuring the qual- ity and relevance of materials science data. This often requires domain expertise, as different materials, such as polymers, metal-organic frameworks, and high-entropy alloys, have dis- tinct physical and chemical properties, methods, and termi- nologies. For instance, classifying superconductors can be complex, combining chemical-based classes like cuprates [11] and iron-based materials [12] with phenomenological cate- gories like heavy fermions [13]. Furthermore, confusion can arise from the polysemy of terms across sub-domains, where the same term might have different meanings. For example, ”TC” may refer to ”Curie Temperature” or ”critical tem- perature” depending on the context. These domain-specific conventions pose significant challenges when building cross- domain datasets.

Problem Statement: The advent of large language models (LLMs) marks a new era in technology, excelling in connect- ing diverse concepts and engaging in complex reasoning [14], [15], [16], [17]. Rule-based approaches, though simpler and faster, are labor-intensive to refine and struggle with gen- eralization. Small language models (SLMs), such as BERT- based models, are more task-specific but require less fine- tuning due to the diverse data used in pre-training. LLMs, with larger context windows (up to 128,000 tokens for GPT- 4-Turbo), offer an advantage in sustaining contextual memory over BERT models, which are limited to 512 tokens [18]. Interaction with LLMs through prompts makes automation more accessible, but their true reasoning capabilities are still under evaluation.

Previous work on information extraction (IE) suggests LLMs are capable of handling general tasks but underper- form in domain-specific areas [19]. Studies comparing LLMs and SLMs for tasks like named entity recognition (NER), relation extraction (RE), and event detection (ED) across various domains show comparable

ISSN 2394 - 7780

results [20], [21], [22]. In materials science, GPT-4 has demonstrated some understand- ing of chemical compounds [23], but its knowledge remains general, lacking the ability to incorporate real-time scientific literature [24]. This study evaluates the potential of LLMs to comprehend, process, and reason with complex materials science information.

Objectives: This work addresses two key questions:

- **Q1**: How effectively can LLMs extract materials science- related data?
- **Q2**: To what extent can LLMs apply reasoning to relate complex concepts?

Approach: We categorize materials science data for novel materials design into two primary elements: material descrip- tions and property expressions. Property data, such as critical temperatures, follow structured formats, often including modi- fiers, values, and units. Material definitions are more domain- specific, requiring detailed descriptions, including composi- tional ratios and processing methods. Identifying materials is challenging due to inconsistent naming conventions used in research.

To address **Q1**, we evaluate LLM performance on NER tasks, focusing on extracting materials and properties from relevant datasets.

NER [25] is crucial in information extraction, aiming to identify and categorize entities such as materials, dopants, conditions, and properties from unstructured text. This task aligns with sequence labeling, where each text token is tagged with a predefined category, crucial for building structured datasets in materials science.

We address Q2 by assessing the capability to establish connections between a predefined set of entities and ex- tract relationships within a given context. Extracting relations between entities is a foundational undertaking in NLP. It entails discerning connections or associations among entities referenced within textual data. For instance, in biomedical research, relationship extraction might involve identifying the association between specific genes and diseases mentioned in scientific literature.

II. EVALUATION

We assess our model's performance using metrics such as Precision, Recall, and F1-score against a baseline established by scores from a BERT-based encoder and a rule-based algorithm from our prior studies [26], [27]. Our models are required to generate outputs in valid JSON format to facilitate structured database extraction (Section III-A1).

Evaluating generative models adds complexity compared to traditional sequence labeling methods used in Named Entity Recognition (NER). While sequence labeling directly matches input and output tokens, generative models may yield structurally different outputs. In cases of material expressions, we introduce a specialized evaluation method that normalizes materials to their chemical formulas and compares individual elements. For instance, while "solar cell" and "solar cells" re- fer to the same concept, "Ca" (Calcium) and "Cr" (Chromium) are distinct, highlighting the need for a precise evaluation approach.

Our Contributions are Summarized as follows:

- Benchmarking LLMs for information extraction, particu- larly NER of materials and properties (addressing Q1).
- Evaluating LLMs for Relation Extraction (RE) in mate- rials science (addressing Q2).
- Proposing a novel evaluation approach for material entity extraction based on "formula matching" through element comparison.

III. METHOD

We utilize three OpenAI LLM models: GPT-3.5-Turbo (gpt-3.5-turbo-0611), GPT-4 (gpt-4), and GPT-4-Turbo (gpt- 4-0611-preview). The evaluation of open-source LLMs is reserved for future work due to their limitations in generating valid JSON outputs (Section III-A1).

Our evaluation employs zero-shot prompting, few-shot prompting, and fine-tuning. Zero-shot prompting assesses a model's generalization to unfamiliar tasks, while few-shot prompting tests adaptation with minimal examples. Fine- tuning involves adjusting a pre-trained model for specific tasks using a smaller dataset.

We selected two datasets: MeasEval [28] for extracting measurements and SuperMat, an annotated dataset on super- conductors [29]. Baseline scores were established using a SciBERT-based encoder and a rule-based RE algorithm [26]. Evaluation metrics, including Precision ($TP/(TP + FP)$), Recall ($TP/(TP + FN)$), and F1-score ($2 \cdot Precision \cdot Recall/(Precision + Recall)$), were computed through pair- wise

ISSN 2394 - 7780

comparisons of predicted and actual entities. The re- sults encompass average F1 scores and their standard de- viations across three runs, with detailed results provided in

Appendix **??**.

A. Named Entity Recognition

The NER task involves identifying entities such as materials and properties. We employed four matching methods, high- lighting the most relevant:

- **Strict**: Exact matching.
- **Soft**: Ratcliff/Obershelp similarity with a threshold of 0.9.
- **Sentence BERT**: Semantic similarity comparison using a threshold of 0.9.
- **Formula Matching**: Our proposed method normalizing materials to chemical formulas and performing element- by-element comparisons.

Prompts for LLM interaction consist of system and user prompts. The fixed system prompt guides the model, while user prompts request specific information about materials and quantities.

For few-shot prompting, we included suggestions based on prior model outputs, which may not be entirely accurate but serve as examples.

1) ***Output Format:*** We required outputs in valid JSON format for several reasons: a) JSON is machine-readable and easily de-serialized. b) The schema can be documented independently of programming languages. c) JSON is an open standard for universal use.

Formatting instructions in user prompts specified the ex- pected JSON schema, as illustrated below:

2) ***Formula Matching:*** Matching materials with generative models is challenging due to potentially differing outputs. Traditional methods have relied on manual evaluations. We propose a novel *formula_matching* method, enabling element- by-element comparisons of normalized formulas derived from our previous work [26].

This method enhances evaluation precision by comparing materials at a granular level. Details of the evaluation and discussion are found in Section IV-B.

B. Relation Extraction

The baseline for RE is established using a rule-based algorithm from our prior work [26], evaluated with Super-Mat. Prompts are designed to group entities based on their relationships, with a strict matching criterion for scoring.

Further details are presented in Section IV-E. remain rele- vant for both system and user prompts, with the task descrip- tion reiterated in each prompt.

1) ***Shuffled vs. Non-Shuffled Evaluation:*** The order of en- tities provided to the Language Model (LLM) can influence evaluation outcomes, particularly when models generate rela- tions sequentially. This can lead to inflated scores that don't accurately reflect the model's relational reasoning capabilities. To mitigate this, we assess relation extraction (RE) using two approaches: *non-shuffled evaluation*, where entities are presented in their original document order, and *shuffled eval- uation*, where entities are randomly rearranged before being input into the prompt.

C. Fine-Tuning Considerations

We fine-tuned the GPT-3.5-Turbo model on the OpenAI platform, which allowed us to create a new model in just a few hours. As of now, fine-tuning for GPT-4 and GPT-4-Turbo is unavailable. All fine-tuned models utilized default parameters set by OpenAI.

Table V summarizes the dataset sizes used for training. For properties extraction, we fine-tuned using the "grobid- quantities dataset" [27] due to the insufficient examples in MeasEval for reliable evaluation.

The primary challenge with the fine-tuned model was generating valid, machine-readable JSON output. While we initially formatted the training data to expect valid JSON, the model struggled to produce it, likely due to a lack of training examples. To address this, we shifted to a pseudo-format that used spaces and line breaks to simplify the model's output. For instance, the expected output for a RE task was structured as follows:

ISSN 2394 - 7780

This technique allowed us to fine-tune the model for con- versational responses, after which we used the base GPT-3.5- Turbo model to convert the outputs into JSON format.

To enhance fine-tuning for the RE task, we introduced variability in the sorting of entity lists (Section III-B). This method maintains dataset size while reducing the risk of the model learning to group entities based solely on their document order, referred to as ”FT.base.” In Section IV-E1, we discuss two additional strategies for preparing the fine- tuning data. The ”FT.document_order” strategy keeps entities in their original document order, which we found detrimental to performance during shuffling evaluations (Section IV-E). The ”FT.augmented” strategy increases dataset size by gen- erating multiple training records with shuffled entity lists for each example in ”FT.base,” roughly doubling the dataset size (Table V). We anticipate that this approach will yield performance comparable to or better than ”FT.base.”

IV. RESULTS AND DISCUSSIONS

In this section, we present and discuss the formula matching and the aggregated results of our evaluations for the LLMs. The completed raw results are available in the Appendix **??**.

A. Limitation of this Study

In this paper, we aim to estimate how well LLMs work in tasks related to materials science. Due to the lack of clean datasets covering the entire materials science domain, we used a dataset that focuses on superconductor material. While our goal is to propose a methodology, we are aware that our results need to be verified empirically in other materials science sub- domains in future works. The following intuitions support our hypothesis: for material NER, we expect that the forms on which materials are presented in other domains would have similar expressions to the ones used in superconductor re- search, considering that chemical formulas, sample names, and commercial names would unlikely be very different between domains. Furthermore, the properties, expressed as measure- ment and physical quantities, are common to all domains; although the statistical distribution could be different, we don’t expect dramatic differences within materials science. On the other hand, RE tasks surely require more datasets that focus both on different domains and different flavours of the same task. As an example, the MatSCIRe [30] dataset, which covers battery-related research, proposes a structure that challenges the relation extraction only between two entities (binary extraction) with the addition of the type of relation which could be inferred by the properties being extracted. In conclusion, we will remand the generalisation for further work.

B. Formula Matching

We evaluated the formula matching to measure two main pieces of information: the gain in the F1-score, and the correctness, as the number of invalid new matches, of the gain. We compared the formula matching with the strict matching because a) it is simple to reproduce and understand visually, and b) the formula matching is built on top of strict matching. We would have more difficulties explaining matches provided by soft matching or SentenceBERT.

We examined the GPT-3.5-Turbo NER extraction (discussed in Section IV-E). 107 out of the 1402 expected records matched correctly using strict matching (P: 22.5%, R: 13.64%, F1: 17.01%). Applying formula matching on the mismatching records, we obtained an additional 176 matches (P: 61.12%, R: 36.00%, F1: 45.31%), for a total gain in F1-score of 28.3 (+266%). For the new 176 records that the formula matching was identifying, we manually examined each pair finding 5 incorrect matches, which corresponds to an error rate of 2.5%. Most of the mismatches in the strict matching caught up by the formula matching were due to missing adjoined information. The LLMs were not able to include information about doping or shape in the response (e.g. hole-doped La 2-x Sr x CuO 4 was not matching with La 2-x Sr x CuO 4).

In other cases, the formula was different by formatting, like: Nd 2-x Ce x CuO 4 and La 2-x Sr x CuO 4. However, the more interesting cases were provided by element or amount substitutions such as: electron-doped infinite-layer superconductors Sr 0.9 La 0.1 Cu 1-x R x O 2 where R = Zn and Ni which was matched Sr0.9La0.1Cu1-xNixO2, or Eu 1-x K x Fe 2 As 2 samples with x = 0.35, 0.45 and 0.5 and Eu 0.5 K 0.5 Fe 2 As 2’. These two cases were particularly complicated to match because they required a deeper understanding of the formula structure.

Among the errors of the formula matching, all of them were provided by the formula which was not correctly parsed, for example in one complicated case with the substrate information: (1-x/2)La 2 O 3 /xSrCO 3 /CuO in molar ratio with x = 0.063, 0.07, 0.09, 0.10, 0.111 and 0.125 which was incorrectly matched with the general La2O3.

C. NER on Properties Extraction

ISSN 2394 - 7780

Suddenly, none of the models outflanked grobid-amounts in zero-shot provoking, as portrayed in Figure 2. This result is astonishing looking at that as a) the statement of properties comes up short on unambiguous space requirement (beside possible varieties in recurrence dissemination), and b) estima- tions of actual amounts are logical common in the broad text corpus used to pre-train the OpenAI models.

In the domain of few-shot provoking (Figure 2), a minimal improvement was noticed exclusively for GPT-4 and GPT- 4-Super, bringing about a F1-score gain going around 2%. In any case, this improvement isn't huge. We speculate that the clues gave to the LLMs might present predisposition. At the point when these clues are mistaken or fragmented, the LLMs battle to direct the age really, influencing the nature of the result results. Essentially, the calibrated model (Figure 2) shows a slight improvement contrasted with zero-shot, barely any shot, and the benchmark. Curiously, in this particular case where both the gauge and adjusted models are prepared and assessed on similar information, the LLM shows a surmised 3% increment in the F1-score.

D. NER on Materials Expressions Extraction

The assessment of material articulations extraction was performed utilizing the parcel of the SuperMat [29] dataset committed to approval, comprising of 32 articles.

In zero-shot provoking (Figure 3), both GPT-4 and GPT- 4-Super accomplished tantamount F1-scores, drifting around 50%. Prominently, all LLMs scored something like 10% lower than the baseline [26]. This divergence is normal, considering that material articulations might include broad arrangements and envelop different snippets of data not effectively conveyed in the brief. Barely any shot inciting (Figure 3) yielded superior outcomes, with GPT-3.5-Super and GPT-4 somewhat unparalleled the pattern. The presentation of clues in the brief to be sure upgrades execution, however, as recently talked about, it appears to unequivocally impact the LLMs, not ready to relieve the effect of invalid clues that might be given. Similarly startling, adjusting didn't beat not many shot inciting. This result proposes that the extra preparation didn't fundamentally upgrade the LLMs' capacity to deal with material articulations.

E. Relation Extraction

The assessment of RE used the total SuperMat dataset, with the outcomes showed in Figure 4, looking at the impacts of rearranging across various models.

GPT-3.5-Super zero-shot and scarcely any shot provoking exhibit a huge contrast among rearranged and non-rearranged assessment (Section III-B1), recommending a successive asso- ciation of substances without explicit logical thinking. Promi- nently, the adjusted GPT-3.5-Super model outflanks the stan- dard by roughly 15% F1-score and doesn't show important contrasts when the assessment is performed under rearranging conditions.

Figure **??** explicitly features the rearranged variant of each model and extraction type. Aside from GPT-3.5-Super, scarcely any shot provoking shows an improvement contrasted with zero-shot provoking, accomplished by consolidating extra models in each brief. GPT-4 and GPT-4-Super likewise show stable outcomes under rearranging conditions, accomplishing a F1-score of around 15-18% lower than calibrated GPT-3.5- Super.

1) ***Data variability for fine-tuning:*** In Section III-C, we de- pict two extra ways of setting up the information for adjusting. As shown in Figure 6, the GPT-3.5-Super model adjusted with the procedure "FT.document_ order" showed a powerlessness to sum up when assessed under rearranging conditions, where the model loses around 30% in F1-score. This proposes that adding entropy (for instance, by rearranging the information) ought to be proceeded as a best practice, which could bring about models with bigger thinking capacities.

At the point when we expanded the size of the dataset utilized in adjusting to practically twofold (Table **??**), the subsequent model didn't work on contrasted with the FT.base. These outcomes affirm that in calibrating, size doesn't make any difference, while information changeability and quality do.

V. CONCLUSION

In this study, we have proposed an evaluation framework for estimating how well LLMs perform compared with SLMs and rule-based tasks related to materials science by focusing on sub-domains such as superconductor research. The findings obtained from our work provide initial guidance applicable to other materials science sub-domains in future research.

To evaluate material extraction comparison, we proposed a novel method to parse and match formula elements by elements through an aggregated parser for materials. This new method provides a more realistic F1 score. Compared with strict matching, we obtained a gain in F1-score from 17% to 45% for GPT3.5-Turbo NER at the price of a minimal error rate (2%).

ISSN 2394 - 7780

We then assessed LLMs on two assignments: NER for materials and properties and RE for connecting them. LLMs fail to meet expectations essentially on NER undertakings than SLMs in material and property extraction (Q1). This finding is especially amazing considering properties since these articulations are not bound to a particular space.

In material extraction, GPT-3.5-Super with calibrating ne- glected to outflank the standard, and similar holds for any model with not many shot provoking. For property extrac- tion, GPT-4 and GPT-4-Super with zero-shot inciting perform comparable to the standard. GPT-3.5-Super with few-shot and calibrating, then again, beats the standard by a minor expansion in focuses. That's what our outcomes recommend, for material articulations, little particular models stay the most dependable decision.

The situation improves for RE (Q2). With two models, barely any shot inciting exhibits a critical improvement over the standard. GPT-4-Super shows upgraded thinking abilities contrasted with GPT-4 and GPT-3.5-Super. GPT-3.5-Super performs ineffectively in both zero-shot and scarcely any shot provoking, showing a significant score decline when elements are rearranged, which lines up with past perceptions. By the by, calibrating yields scores better than the gauge and different models, showing soundness while looking at rearranged and unshuffled assessments.

All in all, to answer Q2, GPT-4 and GPT-4-Super feature powerful thinking abilities for precisely relating ideas and removing relations without calibrating. Be that as it may, calibrating GPT-3.5-Super out yields the best outcomes with a generally little dataset. GPT-4-Super, which costs 33Notwith- standing, for Q1, for removing complex substances, for exam- ple, materials, we find that preparing little particular models stays a more compelling methodology.

REFERENCES

[1] P. Xu, X. Ji, M. Li, and W. Lu, "Small data machine learning in materials science," *npj Computational Materials*, vol. 9, no. 1, p. 42, Mar. 2023.

[2] P. G. Boyd, A. Chidambaram, E. Garc´ıa-D´ıez, C. P. Ireland, T. D. Daff, R. Bounds, A. Gładysiak, P. Schouwink, S. M. Moosavi, M. M. Maroto- Valer *et al.*, "Data-driven design of metal–organic frameworks for wet flue gas co2 capture," *Nature*, vol. 576, no. 7786, pp. 253–256, 2019.

[3] Z. Rao, P.-Y. Tung, R. Xie, Y. Wei, H. Zhang, A. Ferrari, T. Klaver, Ko¨rmann, P. T. Sukumar, A. Kwiatkowski da Silva *et al.*, "Machine learning–enabled high-entropy alloy discovery," *Science*, vol. 378, no. 6615, pp. 78–85, 2022.

[4] A. Zakutayev, N. Wunder, M. Schwarting, J. D. Perkins, R. White, K. Munch, W. Tumas, and C. Phillips, "An open experimental database for exploring inorganic materials," *Scientific data*, vol. 5, no. 1, pp. 1–12, 2018.

[5] T. D. Huan, A. Mannodi-Kanakkithodi, C. Kim, V. Sharma, G. Pilania, and R. Ramprasad, "A polymer dataset for accelerated property predic- tion and design," *Scientific data*, vol. 3, no. 1, pp. 1–10, 2016.

[6] E. O. Pyzer-Knapp, J. W. Pitera, P. W. Staar, S. Takeda, T. Laino, D. P. Sanders, J. Sexton, J. R. Smith, and A. Curioni, "Accelerating materials discovery using artificial intelligence, high performance computing and robotics," *npj Computational Materials*, vol. 8, no. 1, p. 84, 2022.

[7] N. Huber, S. R. Kalidindi, B. Klusemann, and C. J. Cyron, "Machine learning and data mining in materials science," p. 51, 2020.

[8] G. Park and L. Pouchard, "Advances in scientific literature mining for interpreting materials characterization," *Machine Learning: Science and Technology*, vol. 2, no. 4, p. 045007, 2021.

[9] S. Chittam, B. Gokaraju, Z. Xu, J. Sankar, and K. Roy, "Big data mining and classification of intelligent material science data using machine learning," *Applied Sciences*, vol. 11, no. 18, 2021. [Online]. Available: https://www.mdpi.com/2076-3417/11/18/8596

[10] B. Ma, X. Wei, C. Liu, X. Ban, H. Huang, H. Wang, W. Xue, S. Wu, M. Gao, Q. Shen, M. Mukeshimana, A. O. Abuassba, H. Shen, and Y. Su, "Data augmentation in microscopic images for material data mining," *npj Computational Materials*, vol. 6, no. 1, p. 125, 2020.

[11] I. A. Parinov, *Microstructure and properties of high-temperature super- conductors*. Springer Science & Business Media, 2013.

[12] H. Hosono, K. Tanabe, E. Takayama-Muromachi, H. Kageyama, S. Ya- manaka, H. Kumakura, M. Nohara, H. Hiramatsu, and S. Fujitsu, "Exploration of new superconductors and functional materials, and fab- rication of superconducting tapes and wires of iron pnictides," *Science and Technology of Advanced Materials*, 2015.

[13] K. Mydeen, A. Jesche, K. Meier-Kirchner, U. Schwarz, C. Geibel, H. Rosner, and M. Nicklas, "Electron doping of the iron-arsenide superconductor cefeaso controlled by hydrostatic pressure," *Physical Review Letters*, vol. 125, no. 20, p. 207001, 2020.

[14] C. Zhang, C. Zhang, C. Li, Y. Qiao, S. Zheng, S. K. Dam, M. Zhang, J. U. Kim, S. T. Kim, J. Choi *et al.*, "One small step for generative ai, one giant leap for agi: A complete survey on chatgpt in aigc era," *arXiv preprint arXiv:2304.06488*, 2023.

[15] S. Yao, D. Yu, J. Zhao, I. Shafran, T. L. Griffiths, Y. Cao, and K. Narasimhan, "Tree of thoughts: Deliberate problem solving with large language models," *arXiv preprint arXiv:2305.10601*, 2023.

[16] K. Valmeekam, M. Marquez, S. Sreedharan, and S. Kambhampati, "On the planning abilities of large language models–a critical investigation," *arXiv preprint arXiv:2305.15771*, 2023.

[17] S. Sun, Y. Liu, S. Wang, C. Zhu, and M. Iyyer, "Pearl: Prompting large language models to plan and execute actions over long documents," *arXiv preprint arXiv:2305.14564*, 2023.

[18] OpenAI. (2024) Models. https://platform.openai.com/docs/models. [On- line; accessed 04-January-2024].

[19] J. Kocon´, I. Cichecki, O. Kaszyca, M. Kochanek, D. Szydło, J. Baran, J. Bielaniewicz, M. Gruza, A. Janz, K. Kanclerz, A. Kocon´, B. Koptyra, W. Mieleszczenko-Kowszewicz, P. Miłkowski, M. Oleksy, M. Piasecki, Ł. Radlin´ski, K. Wojtasik, S. Woz´niak, and P. Kazienko, "ChatGPT: Jack of all trades, master of none," *Information Fusion*, vol. 99, p. 101861, nov 2023. [Online]. Available: https://doi.org/10.1016%2Fj.inffus.2023.101861

[20] Y. Ma, Y. Cao, Y. Hong, and A. Sun, "Large language model is not a good few-shot information extractor, but a good reranker for hard samples!" *arXiv preprint arXiv:2303.08559*, 2023.

[21] C.-E. Gonza´lez-Gallardo, E. Boros, N. Girdhar, A. Hamdi, J. G. Moreno, and A. Doucet, "Yes but.. can chatgpt identify entities in historical documents?" *arXiv preprint arXiv:2303.17322*, 2023.

[22] M. Moradi, K. Blagec, F. Haberl, and M. Samwald, "Gpt-3 models are poor few-shot learners in the biomedical domain," *arXiv preprint arXiv:2109.02555*, 2021.

[23] K. Hatakeyama-Sato, N. Yamane, Y. Igarashi, Y. Nabae, and T. Hayakawa, "Prompt engineering of gpt-4 for chemical research: what can/cannot be done?" *Science and Technology of Advanced Materials: Methods*, vol. 3, no. 1, p. 2260300, 2023. [Online]. Available: https://doi.org/10.1080/27660400.2023.2260300

[24] K. Hatakeyama-Sato, S. Watanabe, N. Yamane, Y. Igarashi, and K. Oy- aizu, "Using gpt-4 in parameter selection of polymer informatics: improving predictive accuracy amidst data scarcity and 'ugly duck- ling'dilemma," *Digital Discovery*, vol. 2, no. 5, pp. 1548–1557, 2023.

[25] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.

[26] L. Foppiano, P. Castro, P. Suarez, K. Terashima, Y. Takano, and M. Ishii, "Automatic extraction of materials and properties from superconductors scientific literature," *Science and Technology of Advanced Materials Methods*, vol. 3, 2023.

[27] L. Foppiano, L. Romary, M. Ishii, and M. Tanifuji, "Automatic identification and normalisation of physical measurements in scientific literature," in *Proceedings of the ACM Symposium on Document Engineering 2019*, ser. DocEng '19. New York, NY, USA: Association for Computing Machinery, 2019. [Online]. Available: https://doi.org/10.1145/3342558.3345411

[28] C. Harper, J. Cox, C. Kohler, A. Scerri, R. Daniel Jr., and P. Groth, "SemEval-2021 task 8: MeasEval – extracting counts and measurements and their related contexts," in *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, A. Palmer, N. Schneider, N. Schluter, G. Emerson, A. Herbelot, and X. Zhu, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 306– 316. [Online]. Available: https://aclanthology.org/2021.semeval-1.38

ISSN 2394 - 7780

[29] L. Foppiano, T. Dieb, A. Suzuki, P. Castro, S. Iwasaki, A. Uzuki, M. Echevarria, Y. Meng, K. Terashima, L. Romary, Y. Takano, and M. Ishii, "Supermat: construction of a linked annotated dataset from superconductors-related publications," *Science and Technology of Ad- vanced Materials Methods*, vol. 1, pp. 34–44, 2021.

[30] A. Mullick, A. Ghosh, G. S. Chaitanya, S. Ghui, T. Nayak, S.-C. Lee, S. Bhattacharjee, and P. Goyal, "Matscire: Leveraging pointer networks to automate entity and relation extraction for material science knowledge-base construction," *Computational Materials Science*, vol. 233, p. 112659, 2024.

FIGURES & TABLES

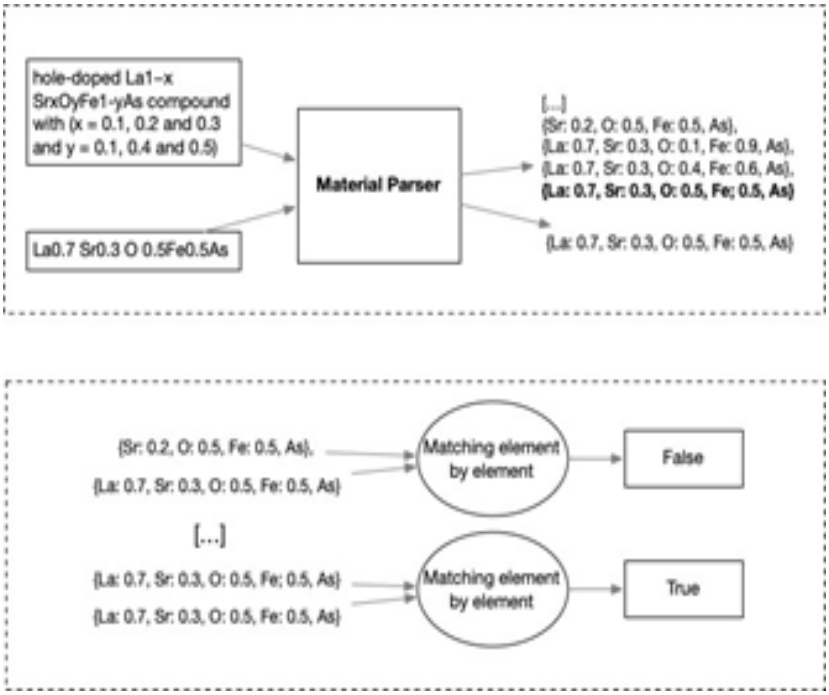

Fig. 1: Two materials that appear to have a very different composition are, in reality, overlapping.

Task	Preparation strategy	Dataset	# Training	#
NER	N/A	SuperMat	1639	703
NER	N/A	grobid-quantities dataset	485	208
RE	FT.base/FT.document	SuperMat	344	148
RE	FT.augmented	SuperMat	695	299

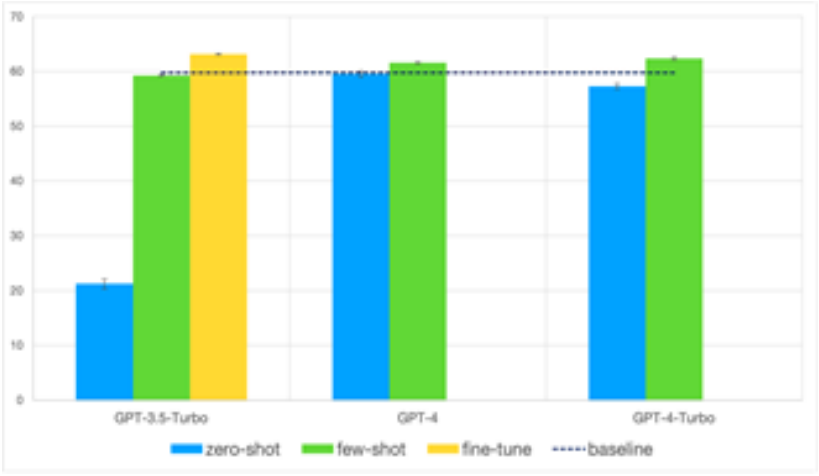

Fig. 2: Comparison scores for properties extraction using NER.

ISSN 2394 - 7780

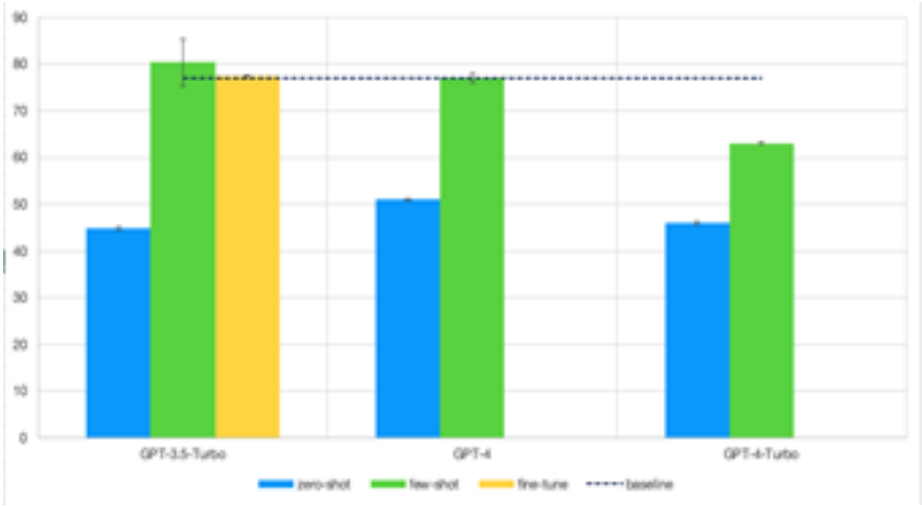

Fig. 3: Comparison scores for material extraction using NER.

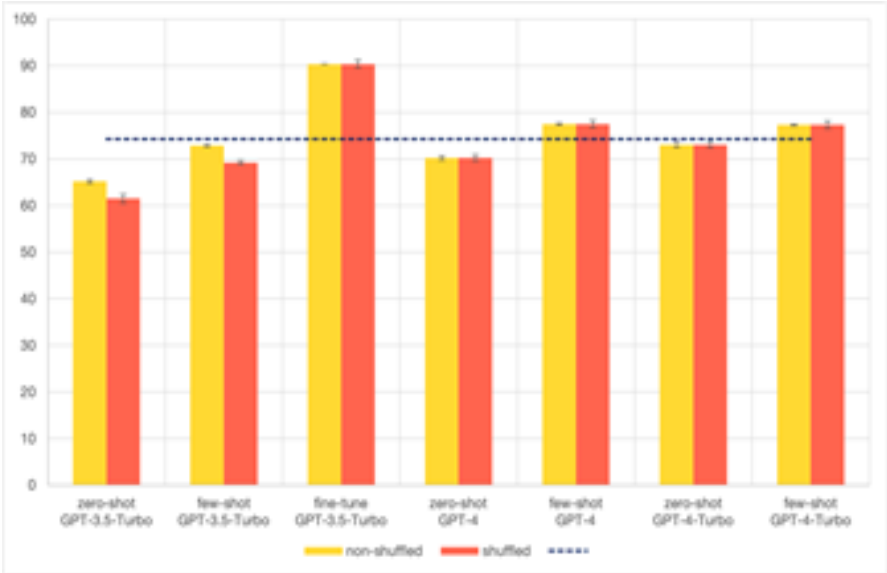

Fig 4: Scores of the shuffled extraction using zero-shot prompting, few-shot prompting and the fine-tuned model for RE on materials and properties.

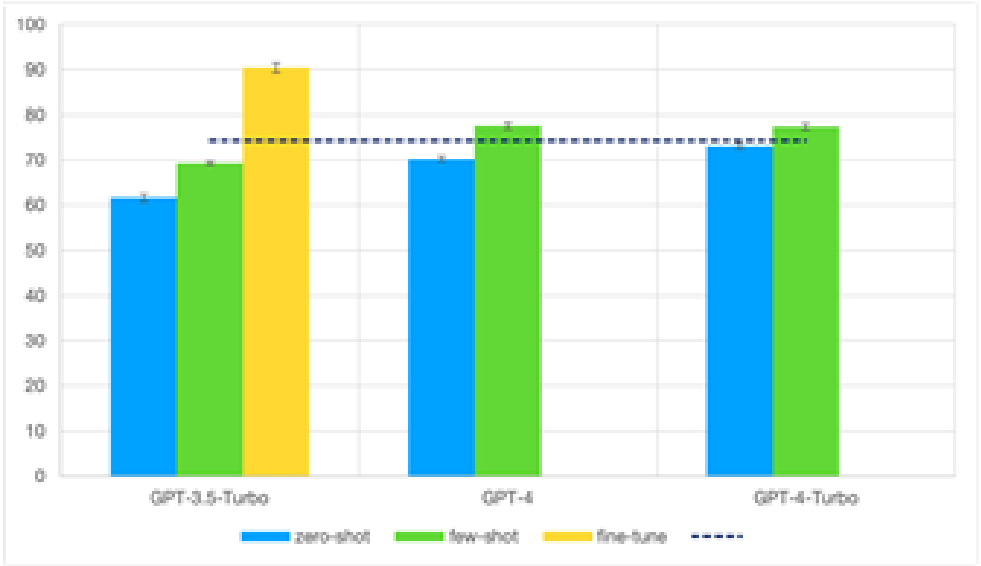

Fig. 5: Overall evauation.

ISSN 2394 - 7780

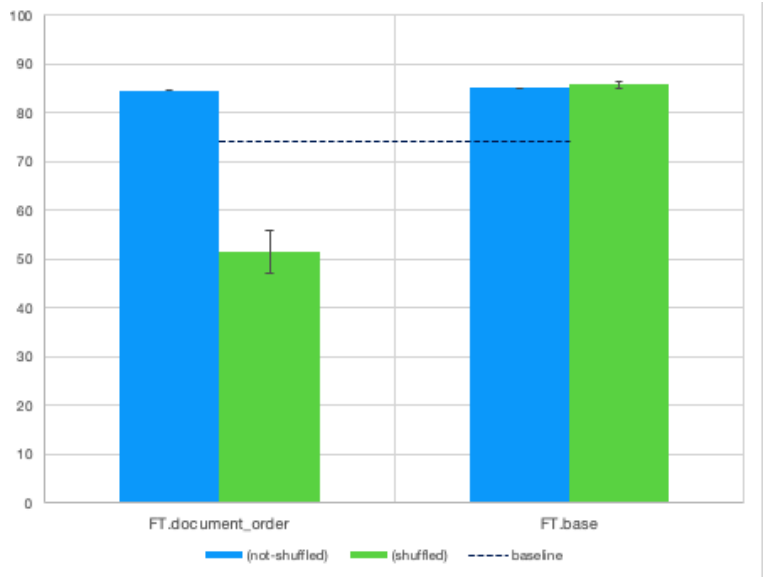

Fig. 6: Evaluation of the impact of data variability in fine-tuning GPT-3.5- Turbo.