



PDF Download
3746252.3761359.pdf
02 February 2026
Total Citations: 0
Total Downloads: 177

Latest updates: <https://dl.acm.org/doi/10.1145/3746252.3761359>

RESEARCH-ARTICLE

Towards Fully-Automated Materials Discovery via Large-Scale Synthesis Dataset and Expert-Level LLM-as-a-Judge

HEEGYU KIM, Ajou University, Suwon, Gyeonggi-do, South Korea

TAEYANG JEON, Ajou University, Suwon, Gyeonggi-do, South Korea

SEUNGTAEK CHOI, Hankuk University of Foreign Studies, Seoul, South Korea

JI-HOON HONG, Ajou University, Suwon, Gyeonggi-do, South Korea

DONG-WON JEON, Ajou University, Suwon, Gyeonggi-do, South Korea

GA-YEON BAEK, Hanyang University, Seoul, South Korea

[View all](#)

Open Access Support provided by:

[Ajou University](#)

[Hanyang University](#)

[Hankuk University of Foreign Studies](#)

Published: 10 November 2025

[Citation in BibTeX format](#)

CIKM '25: The 34th ACM International Conference on Information and Knowledge Management
November 10 - 14, 2025
Seoul, Republic of Korea

Conference Sponsors:
[SIGWEB](#)
[SIGIR](#)

Towards Fully-Automated Materials Discovery via Large-Scale Synthesis Dataset and Expert-Level LLM-as-a-Judge

Heegyu Kim*
Ajou University
Suwon, Republic of Korea

Ji Hoon Hong
Ajou University
Suwon, Republic of Korea

Gyeong-Won Kwak
Hanyang University
Seoul, Republic of Korea

Chihoon Lee
Hanyang University
Seoul, Republic of Korea

Jin-Seong Park
Hanyang University
Seoul, Republic of Korea

Taeyang Jeon*
Ajou University
Suwon, Republic of Korea

Dong Won Jeon
Ajou University
Suwon, Republic of Korea

Dong-Hee Lee
Hanyang University
Seoul, Republic of Korea

Yoon-Seo Kim
Hanyang University
Seoul, Republic of Korea

Sung Beom Cho
Ajou University
Suwon, Republic of Korea

Seungtaek Choi
Hankuk University of Foreign Studies
Seoul, Republic of Korea

Ga-Yeon Baek
Hanyang University
Seoul, Republic of Korea

Jisu Bae
Hanyang University
Seoul, Republic of Korea

Seon-Jin Choi
Hanyang University
Seoul, Republic of Korea

Hyunsouk Cho[†]
hyunsouk@ajou.ac.kr
Ajou University
Suwon, Republic of Korea

ABSTRACT

Materials synthesis remains a critical bottleneck in developing innovations for energy storage, catalysis, electronics, and biomedical devices. Current synthesis design relies heavily on empirical trial-and-error methods guided by expert intuition, limiting the pace of materials discovery. To address this challenge, we present AlchemyBench, a comprehensive benchmark built upon a curated dataset of 17,667 expert-verified synthesis recipes from open-access literature.

AlchemyBench provides an end-to-end framework that supports research in large language models (LLMs) applied to materials synthesis prediction. The benchmark encompasses four key tasks: raw materials and equipment prediction, synthesis procedure generation, and characterization outcome forecasting. To enable scalable evaluation, we propose an LLM-as-a-Judge framework that leverages large language models for automated assessment, demonstrating strong agreement with expert evaluations (e.g., Pearson's $r = 0.80$, Spearman's $\rho = 0.78$).

Our experimental results reveal that reasoning-focused models (Claude 3.7, GPT-4o) achieve scores around 4.0 on well-documented oxide and organic synthesis targets, but performance drops by approximately 0.3 points on electrochemical workflows. Fine-tuning on AlchemyBench data enables a 7B-parameter open-source model

to surpass generic baselines trained on 1M samples, while retrieval-augmented generation provides an additional +0.20 improvement when supplied with five high-similarity contexts.

AlchemyBench addresses a critical gap in the field by providing the first comprehensive, legally redistributable benchmark for automated materials synthesis prediction. Our contributions establish a foundation for exploring LLM capabilities in predicting and guiding materials synthesis, ultimately accelerating experimental design and innovation in materials science.

CCS Concepts

• **Applied computing** → **Chemistry**; • **Computing methodologies** → **Language resources**; **Natural language generation**; **Information extraction**.

Keywords

Large Language Model ; LLM-as-a-Judge ; Dataset ; Benchmark ; Human Evaluation ; Materials Science

ACM Reference Format:

Heegyu Kim, Taeyang Jeon, Seungtaek Choi, Ji Hoon Hong, Dong Won Jeon, Ga-Yeon Baek, Gyeong-Won Kwak, Dong-Hee Lee, Jisu Bae, Chihoon Lee, Yoon-Seo Kim, Seon-Jin Choi, Jin-Seong Park, Sung Beom Cho, and Hyunsouk Cho. 2025. Towards Fully-Automated Materials Discovery via Large-Scale Synthesis Dataset and Expert-Level LLM-as-a-Judge. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25)*, November 10–14, 2025, Seoul, Republic of Korea. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3746252.3761359>

*Both authors contributed equally to this research.

[†]Corresponding Author.



This work is licensed under a Creative Commons Attribution 4.0 International License. *CIKM '25, Seoul, Republic of Korea*

© 2025 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2040-6/2025/11

<https://doi.org/10.1145/3746252.3761359>

1 INTRODUCTION

Materials synthesis underpins advances in energy storage, catalysis, electronics, and biomedical devices [30]. Despite its importance, synthesis processes remain largely empirical, relying on trial-and-error approaches guided by expert intuition [25]. Designing a successful recipe is challenging because even minor variations in temperature, reagent order, or concentration can drastically change the outcome; thus, a good synthesis procedure must be both coherent and reproducible. This inefficiency highlights the need for systematic, data-driven approaches to predict synthesis workflows and optimize experimental design [14].

Recent progress in machine learning and large language models has opened new avenues for extracting and generating synthesis procedures from unstructured scientific literature [10, 38]. However, practical adoption is hampered by several challenges. Existing datasets are often small, domain-specific, and noisy, limiting model generalizability. [39, 42] Moreover, the absence of comprehensive benchmarks makes it difficult to assess the performance of synthesis prediction methods, while expert evaluations remain too costly and time-consuming for large-scale use.

To address these challenges, we introduce **Open Materials Guide (OMG)**, a dataset comprising 17K high-quality, expert-verified synthesis recipes curated from open-access literature. This dataset is the foundation for our benchmark, **AlchemyBench**, which evaluates synthesis prediction across multiple facets from inferring raw materials and recommending appropriate synthesis equipment to generating detailed procedural steps and forecasting suitable characterization techniques.

Additionally, we investigate an LLM-as-a-Judge framework to automate the evaluation process. In a human baseline study, we recruited domain experts (master’s level and above) to rate 20 synthesis recipes a task that consumed approximately 460 minutes of cumulative expert time. Our systematic comparisons reveal that LLM-based assessments achieve a high degree of statistical agreement with expert judgments (e.g., Pearson’s $r = 0.80$, Spearman’s $\rho = 0.78$), while imposing negligible human labor. These findings underscore the potential of LLMs to serve as scalable, cost-effective evaluators, rapidly replacing expensive, time-intensive expert review in materials synthesis.

Our work makes the following key contributions:

- **Open Materials Guide (OMG)**, a large-scale, high-quality dataset of 17,667 expert-verified materials synthesis recipes curated from open-access literature. It spans diverse synthesis methods and element combinations, overcoming significant limitations of prior datasets in completeness, reproducibility, and legal reusability.
- **AlchemyBench**, the first end-to-end benchmark for evaluating synthesis prediction tasks including raw material selection, equipment inference, procedure generation, and characterization prediction, powered by a validated LLM-as-a-Judge framework that aligns closely with expert ratings and enables scalable, automated evaluation.
- **Comprehensive empirical evaluation** of leading LLMs under multiple setups (zero-shot, fine-tuning, RAG), providing insights into task difficulty (e.g., high-impact vs. standard recipes), synthesis method variance, and the effect of element

frequency on prediction accuracy. Our findings highlight the promise and limitations of current LLMs in fully automated materials synthesis.

To enhance reproducibility and accessibility, we release the dataset and code as an open-source resource for the research community¹.

2 RELATED WORKS

Materials Synthesis Datasets: Existing materials synthesis datasets, such as those focusing on solid-state [19] and solution-based [40] methods, have provided valuable resources for machine learning applications. However, these datasets often suffer from issues of incompleteness and low quality, with many synthesis procedures lacking critical parameters necessary for reproducibility or predictive modeling. This challenge is further underscored by Sun and David [39], who critically evaluate text-mined solid-state synthesis datasets and find that many recipes either lack key operations or contain only one uninformative step, limiting their usefulness for guiding predictive synthesis workflows.

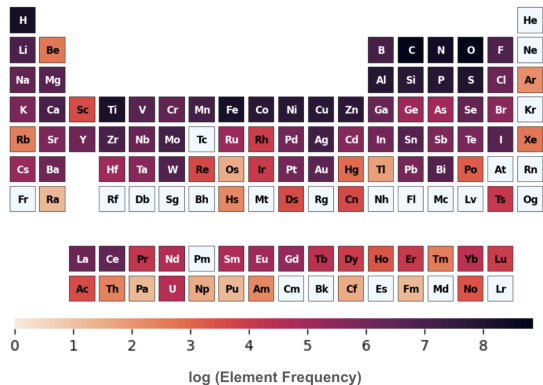
To address these challenges, MS-MENTIONS [29] introduced a large-scale, expert-annotated corpus of 595 materials synthesis procedures, with fine-grained entity labels across 15 mention types, including operations, materials, apparatuses, and conditions. While this resource enables strong performance on mention-level extraction tasks, its scope is limited to local entity recognition rather than supporting full synthesis workflow modeling or generative tasks. Furthermore, manual expert annotation at this scale remains costly and may not always support open or scalable reuse, highlighting the need for more scalable solutions for large-scale benchmark construction.

LLM-Based Generation for Materials Science: Large language models (LLMs) have shown promise in accelerating materials discovery by automating hypothesis generation [20], property prediction [7], and evaluation [27]. However, prior efforts have focused on narrow tasks such as entity recognition or reaction classification, often requiring costly manual annotation or extensive fine-tuning. In contrast, our work leverages LLMs as both generators and evaluators of full synthesis workflows, introducing a large-scale benchmark that captures real-world procedural complexity and enables scalable, expert-level evaluation.

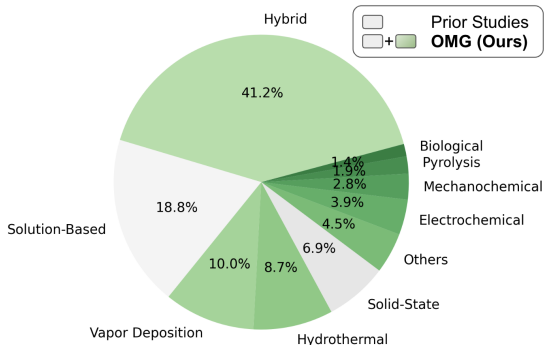
Evaluation for Materials Science: LLMs have shown significant potential in materials synthesis research, outperforming traditional machine learning models in tasks such as precursor selection and synthesizability prediction [18, 21]. Tools like LLMat-Design [17] autonomously generate and evaluate hypotheses for material design, while benchmarks like ALDBench [44] assess LLMs’ ability to address domain-specific questions. However, their application in evaluating synthesis quality remains underexplored, with limited systematic comparisons between LLM-based evaluations and human expert judgments in materials science. This highlights the need for further research to establish LLMs as reliable evaluators aligned with expert assessments.

Our work fills this gap while enabling orders-of-magnitude cheaper, reproducible assessment of synthesis-prediction models.

¹<https://github.com/HeegyuKim/AlchemyBench>



(a) The periodic table of logarithmic frequency of elements.



(b) The distribution of synthesis techniques.

Figure 1: The periodic table (left) demonstrates that OMG covers diverse elements used in target materials, with darker colors indicating higher usage frequencies. A pie chart (right) illustrates the diversity of synthesis methods, highlighting the contributions of prior studies (white) and our dataset (white + green).

3 DATA COLLECTION AND PREPARATION

3.1 Motivation

Previous large-scale datasets for extracting synthesis procedures from materials science literature have faced several critical challenges [19, 40]. The most significant limitation involves common extraction errors, such as missing reagent concentrations, incorrect reaction temperatures, and misordered procedural steps, which have rendered many outputs unreliable for downstream synthesis prediction [39]. We analyzed existing datasets and revealed that over 92% of records in Kononova et al. and 98% in Wang et al. lacked essential synthesis parameters (e.g., heating temperature, duration, mixing media). Additionally, these datasets are narrowly focused on a few synthesis techniques (such as solid-state and solution-based). At the same time, real-world materials innovation employs a broader range of specialized techniques [43]. Finally, copyright restrictions from commercial journals have limited the legal redistribution of textual synthesis procedures [4].

To overcome these limitations, we propose **OMG** with three innovations: an LLM-driven parsing approach that improves extraction accuracy, a systematic collection covering more than ten distinct synthesis techniques (including vapor deposition, hydrothermal, and hybrid material systems), and the exclusive use of open-access publications to enable legal distribution of the dataset.

3.2 Dataset Construction

Our pipeline begins by retrieving 28,685 open-access articles from a pool of 400K search results using the Semantic Scholar API with 60 domain-specific search terms (e.g., “solid-state sintering process”, “metal organic CVD”) recommended by domain experts. We convert PDFs to structured Markdown using PyMuPDFLLM [3] and then employ GPT-4o in a multi-stage annotation process. First, articles are categorized based on whether they contain synthesis protocols, target materials, synthesis techniques, and applications. For articles confirmed to include synthesis procedures, the text is segmented into five key components, as illustrated in Figure 2:

- **X**: A summary of the target material, synthesis method, and application.
- **Y_M**: Raw materials, including quantitative details.
- **Y_E**: Equipment specifications.
- **Y_P**: Step-by-step procedural instructions.
- **Y_C**: Characterization methods and results.

This systematic extraction yielded a dataset of 17,667 high-quality recipes (approximately a 62% yield) covering 10 diverse synthesis methods. Figure 1 demonstrates our dataset’s broad coverage of materials systems and synthesis techniques. We used 60 domain-specific search terms selected by domain experts, and designed GPT-4o [16] prompts to extract structured synthesis components.

3.3 Quality Verification

To ensure the accuracy of our automatically extracted recipes, we assembled a panel of eight domain experts from three institutions. The experts manually reviewed a representative sample of ten recipes, evaluating them based on the following criteria:

- **Completeness**: Capturing the full scope of the reported recipe (**X**, **Y_M**, **Y_E**, **Y_P**, and **Y_C**).
- **Correctness**: Extracting critical details such as temperature values and reagent amounts accurately.
- **Coherence**: Retaining a logical, consistent narrative without contradictions or abrupt transitions.

Table 1 presents our expert evaluation results using a five-point Likert scale (1 = poor, 5 = excellent). To measure expert agreement, we computed the **Intraclass Correlation Coefficient (ICC)** [37], specifically adopting the ICC (3,k) form of a two-way mixed-effects model that evaluates absolute agreement among multiple raters on each item. While alternative metrics like Cohen’s Kappa are commonly used for inter-rater agreement, ICC (3,k) is more appropriate for continuous scores from multiple experts than categorical labels, reflecting consistency and absolute agreement. The extracted data exhibited high mean scores, but inter-rater reliability varied across criteria, particularly for articles with well-structured experimental sections.

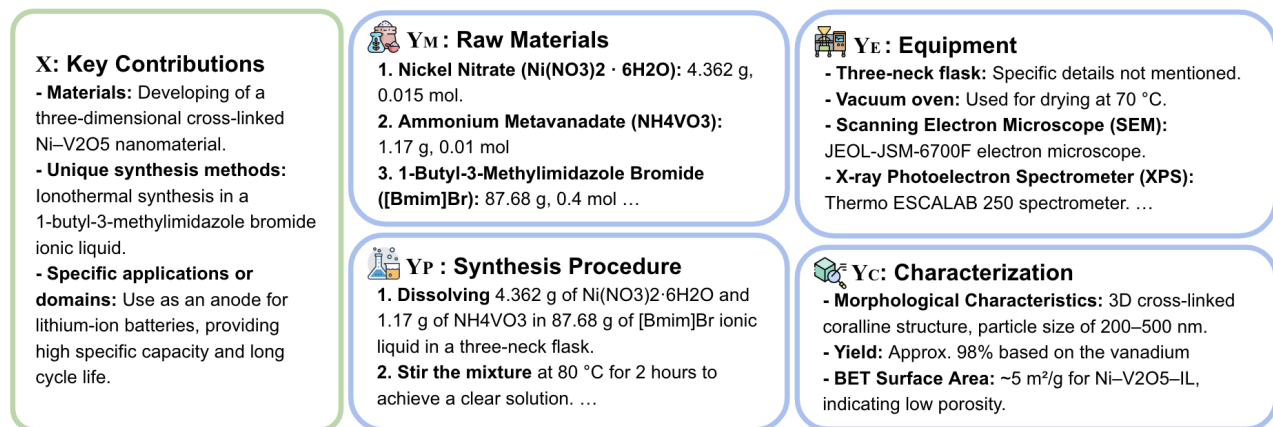


Figure 2: An example of an extracted recipe from Zhao et al. demonstrates structured annotation of materials, equipment, procedures, and characterization methods.

Table 1: Data verification by eight domain experts.

Criteria	Mean σ	ICC (3,k) _{p-value}
Completeness	4.2 0.81	0.695 0.00
Correctness	4.7 0.58	0.258 0.23
Coherence	4.8 0.46	0.429 0.10

Completeness showed moderate agreement (ICC = 0.695), while correctness (ICC = 0.258) and coherence (ICC = 0.429) had lower agreement due to variations in naming conventions and missing characterization details. Although the completeness score (4.2/5.0) was slightly lower than those for correctness (4.7/5.0) and coherence (4.8/5.0), correctness and coherence exhibited lower inter-rater reliability (ICC = 0.258 and 0.429, respectively), suggesting inconsistencies in how evaluators interpreted minor details. Variability in scores for correctness and coherence arose from differences in how evaluators weighted minor inconsistencies, such as variations in equipment naming or missing characterization information. Some considered these negligible, while others applied stricter criteria, underscoring the need for refined annotation guidelines.

While manual verification confirms the effectiveness of our extraction process, it cannot fully ensure consistent performance across the diverse range of synthesis procedures. In the following section (Section 4), we present a structured evaluation framework for tasks such as raw materials and equipment inference, procedure generation, and characterization outcome forecasting.

4 ALCHEMYBENCH

We present **AlchemyBench**, a benchmark for evaluating materials synthesis prediction models. This framework addresses key challenges in synthesis recipe evaluation through structured tasks, expert-aligned metrics, and scalable assessment strategies.

4.1 Motivation

Evaluating synthesis predictions presents several fundamental challenges:





- **Lack of Benchmarks:** No standardized evaluation framework exists, making it challenging to compare synthesis models systematically. Prior datasets lack critical synthesis parameters and structured ground truth labels, making meaningful comparisons difficult.
- **Limitations of Traditional Metrics:** Traditional metrics, such as BLEU [34] and ROUGE [23] prioritize lexical overlap but fail to capture the procedural correctness of synthesis recipes. Na et al. introduced the Jaccard score to measure set overlap in synthesis procedures, yet it lacks sensitivity to sequential dependencies critical in procedural texts. BERTScore [45] improves contextual similarity measurement but struggles with domain-specific dependencies unique to materials synthesis. Moreover, these metrics do not account for experimental feasibility, limiting their applicability in real-world synthesis.
- **High Cost of Human Evaluation:** Expert-based assessments require significant time and resources, averaging 460 minutes for 20 predictions in our experiment. This cost makes large-scale benchmarking impractical, requiring an automated evaluation system.
- **Scalability Requirements:** Large-scale benchmarking necessitates an automated yet reliable evaluation system, which LLMs can provide [13]. However, prior attempts to use LLMs for evaluation lacked systematic validation against human expert assessments in materials science, raising concerns about reliability.

4.2 Task Definition

AlchemyBench simulates real-world synthesis workflows, where models must predict the following components given input X (target material, synthesis method, application domain):

- **P_M:** Raw materials (e.g., reagents, solvents) with quantities.
- **P_E:** Required equipment (e.g., furnace, autoclave).
- **P_P:** Synthesis procedures (e.g., reaction steps, temperatures).
- **P_C:** Characterization methods and expected outcomes.

Table 2: Seven evaluation criteria used to evaluate synthesis recipes, categorized into materials, equipment, procedure, characterization, and overall score. Each criterion is rated on a 1–5 scale to reflect the quality and practicality of the predicted recipes.

Category	Criteria	Description
 Materials	Appropriateness	Are the selected materials suitable for the target synthesis?
 Equipment	Appropriateness	Is the selected equipment suitable?
 Procedure	Completeness	Is the procedure well-organized and logically structured?
	Similarity	How closely does it match the ground truth procedure?
	Feasibility	Can this procedure be realistically executed in a lab?
 Characterization	Appropriateness	Are the methods and metrics suitable for validating the success of the synthesized material?
	Similarity	How well do predicted properties match actual results?
Overall Score	-	Average score considering the recipe’s overall quality and practicality.

Predictions $P_X = \{P_M, P_E, P_P, P_C\}$ are evaluated against ground truth $Y_X = \{Y_M, Y_E, Y_P, Y_C\}$ using the LLM-as-a-Judge framework [47]. Unlike prior benchmarks that rely on lexical similarity, **AlchemyBench** assesses procedural correctness and experimental feasibility. The evaluation criteria are described in Table 2.

The scoring function $\text{Score}(P_X, Y_X)$ is computed as $\frac{\sum_{i=1}^{N_C} C_i}{N_C}$, where C_i represents the score for criterion i , and N_C is the total number of evaluation criteria. These criteria were developed in collaboration with domain experts to ensure alignment with real-world synthesis evaluation.

4.3 Dataset Splits and Distribution

We divided **OMG** to three splits to ensure robust evaluation:

- **Training Set:** 16,026 articles published before 2024.
- **Test - Standard Impact:** 1,472 articles (2024 and beyond) from journals with Impact Factor (IF) < 10.
- **Test - High Impact:** 169 articles (2024 and beyond) from journals with IF ≥ 10.

The **temporal split** ensures that models are evaluated on *unseen future research*, mitigating data contamination. Additionally, stratification by **journal impact** allows assessment of a model’s ability to process high-impact findings, often introducing novel and complex synthesis techniques. This split design evaluates both *generalizability* and the ability to meet the rigorous standards of top-tier journals.

5 LLM AS A JUDGE

A reliable evaluation framework is essential for benchmarking synthesis prediction models. This section examines the alignment between LLM-based and human expert judgments, evaluating inter-rater agreement and assessing the effectiveness of LLMs as automated evaluators.

5.1 Evaluation Metrics

We employ two metrics to assess the reliability of our LLM-as-a-Judge. For comparison, we also report results from traditional automatic evaluation methods: BLEU, ROUGE-L, and BERTScore.

Pearson Correlation Coefficient measures how closely LLM scores align with expert ratings on a continuous scale, capturing linear relationships. Finally, the **Spearman’s Rank Correlation** assesses rank-order consistency, beneficial when the relative ranking of recipes is more informative than absolute scores.

5.2 Human Expert Evaluation Setup

Before evaluating whether the reliability of **AlchemyBench** assessment aligns with expert evaluations, we enlisted eight materials science researchers from three institutions to establish a reliable ground truth. Each evaluator had prior experience in experimental synthesis and was selected based on their publication record and domain expertise. Experts independently assessed model-generated recipes using seven criteria (Table 2) on a 1–5 scale. To ensure high-quality assessments, we also collected self-reported confidence scores for each expert.

Based on these scores, we define two expert groups used in subsequent analyses. The **High-Confidence Subgroup** ($n = 3$) consists of the three experts who reported the highest average confidence levels during annotation. The **Full Expert Panels** ($n = 8$) include all eight annotators, encompassing the High-Confidence Subgroup. Hereafter, we refer to these groups as the **High group** and **Full group**, respectively. These groupings are used throughout Section 5 to compare inter-rater reliability and alignment with LLM evaluations.

Ten representative synthesis workflows were selected by a senior materials scientists to ensure diversity, considering factors such as the evaluators’ specialization, material variety, and the synthesis difficulty to avoid overly complex recipes. This selection includes a mix of diverse recipes and core, fundamental synthesis protocols. This small-scale but highly important selection of workflows was used to generate 20 unique predictions, which were generated by two models (GPT-4o-mini and o1-mini), then evaluated by both human experts and LLM judges.

5.3 Inter-Expert Agreement Analysis

To analyze the effect of annotator confidence on evaluation reliability, we compare the **High group** against the **Full group** to examine differences in inter-rater agreement. The comparison between the High Confidence group and the Full group in Table 3

Table 3: ICC (3,k) for each evaluation criterion. “High (n=3)” denotes a High-Confidence Subgroup and “Full (n=8)” denotes Full Expert Panels. Subscripts indicate p -values.

Criteria	High (n=3)	Full (n=8)
Material Appropriateness	0.61 _{0.01}	0.80 _{0.00}
Equipment Appropriateness	0.63 _{0.00}	0.63 _{0.00}
Procedure Completeness	0.46 _{0.05}	0.23 _{0.19}
Procedure Similarity	0.34 _{0.14}	0.13 _{0.31}
Procedure Feasibility	0.70 _{0.00}	−0.58 _{0.88}
Characterization Appropriateness	0.45 _{0.06}	0.78 _{0.00}
Characterization Similarity	0.37 _{0.11}	0.45 _{0.03}
Overall Score (Average)	0.75 _{0.00}	0.68 _{0.00}

highlights key differences in inter-rater reliability. The Full group achieves higher ICC values for *Material Appropriateness* (0.80) and *Characterization Appropriateness* (0.78) compared to the High group (0.61 and 0.45, respectively), indicating better consensus among the broader panel for these criteria. However, the High group shows significantly stronger agreement on *Procedure Feasibility* (ICC = 0.70) than the Full group, which exhibits a negative ICC value (−0.58), suggesting inconsistencies in feasibility evaluations within the larger group. Both groups display similar reliability for *Equipment Appropriateness* (ICC = 0.63). Overall, while larger panels may enhance agreement on straightforward criteria, smaller high-confidence subgroups provide more consistent evaluations for complex aspects like procedural feasibility.

5.4 LLM-Expert Agreement Analysis

Table 4 presents a comparison between model-generated scores and expert evaluations, divided into the High-Confidence Subgroup ($n = 3$) and the Full Expert Panels ($n = 8$). Across both groups, traditional lexical metrics such as BLEU, ROUGE-L, and BERTScore show weak or negative correlations with expert ratings and highly significant t -values (e.g., $t = -70.87$ for BLEU), underscoring their misalignment with human judgment. In contrast, most LLM-based evaluators—especially GPT-4o Aug and GPT-4o Nov—exhibit substantial and statistically significant agreement with expert scores.

To identify which specific criteria drive these alignment patterns, Table 5 reports Pearson and Spearman correlations by criterion. The High group shows more substantial alignment with LLMs on factual and objective dimensions such as *Material Appropriateness* ($r = 0.59$, $p = 0.01$) and *Characterization Similarity* ($r = 0.45$, $p = 0.05$). In contrast, the Full group exhibits greater correlation in more interpretive aspects like *Procedure Similarity* ($r = 0.56$, $p = 0.01$), likely reflecting broader evaluative perspectives.

Some criteria remain inherently difficult to assess consistently. For instance, *Procedure Feasibility* and *Equipment Appropriateness* yield negligible correlation in all settings (e.g., Feasibility: $r = -0.04$, $p = 0.86$), mirroring the low inter-rater agreement among experts (see Table 3). These discrepancies likely reflect the intrinsic subjectivity of such criteria, which inherently demand considerable domain judgment and contextual interpretation, particularly in

evaluating feasibility and equipment decisions under novel synthesis conditions. Rather than signaling noise, these inconsistencies highlight the task’s complexity and the challenges of codifying expert-level reasoning.

This preliminary validation, conducted with a small sample size ($n = 20$), shows that our LLM-as-a-Judge framework demonstrates a satisfactory overall alignment with expert scores, with a composite correlation rising above $r = 0.70$. However, the low inter-rater agreement on subjective criteria like *Procedure Feasibility* and *Equipment Appropriateness* (see Table 3) underscores the inherent difficulty in consistently evaluating these aspects. These results highlight the need for refined evaluation guidelines in future research to address the variability in expert judgment.

6 EXPERIMENTS

We conducted a comprehensive evaluation of large language models (LLMs) for materials synthesis recipe prediction using the **AlchemymBench**. Our experiments are designed to answer key research questions about model performance, robustness, and the relationship between data frequency and predictive accuracy.

6.1 Experiment Setup

To comprehensively evaluate the models, we conducted experiments with the following setup:

Base LLMs: We evaluated ten LLMs, including reasoning-based models, including o3-mini [33], Claude 3.7 Sonnet [1], DeepSeek R1 [9], and general-purpose models, including Gemini 2.0 Flash [11], Qwen 2.5 72B Instruct [35], Llama 3.3 70B Instruct [12], and GPT-4o variants [16]. The knowledge cutoff for OpenAI’s models is set at October 2023, while Llama-3.3 is constrained to December 2023, thereby minimizing the risk of data contamination in our test sets, which encompass data from 2024 and beyond². Other LLMs have knowledge cutoffs beyond 2024 or do not disclose their knowledge cutoff, indicating potential risks of data contamination. We prompt the LLM with a fixed one-shot example from our train set to predict all components (P_X). We set the temperature to zero and max_tokens to 4,096 for general LLMs and 16,384 for reasoning LLMs.

Evaluation Framework: Each model generated synthesis recipes for both the *High Impact set* and *Standard Impact set*. Recipes were evaluated using our LLM-as-a-Judge method based on GPT-4o-Aug. The evaluation criteria focused on *Material Appropriateness*, *Procedural Similarity*, and overall recipe quality.

6.2 RQ1: Can LLMs Generate Material Synthesis Recipes Across Different Difficulty Levels?

Figure 3 demonstrates that reasoning-centric models such as Claude 3.7 Sonnet (think) achieved the highest overall scores, while open-source models like Llama 3.3 70B Instruct lagged behind. Performance consistently declined on the *High Impact set* across all models. Table 6 details the difference in evaluation criteria between the full test set and the *High Impact set* subset. The most pronounced drops were observed in *Materials Appropriateness* (−0.310, −7.98%) and *Procedural Similarity* (−0.209, −7.53%), indicating that LLMs

²OpenAI models information: <https://platform.openai.com/docs/models>.

Table 4: Agreement between LLM judges and expert evaluations: left panel shows alignment with High-Confidence Experts subgroup ($n = 3$), right panel with Full Expert Panels ($n = 8$). Subscripts indicate p -values.

Judge Model	High-Confidence Experts ($n=3$)			Full Expert Panels ($n=8$)		
	Pearson r	Spearman ρ	t-test	Pearson r	Spearman ρ	t-test
BLEU [34]	-0.04 _{0.79}	-0.03 _{0.85}	-40.02 _{0.00}	-0.01 _{0.97}	0.01 _{0.94}	-70.87 _{0.00}
ROUGE-L [23]	-0.08 _{0.63}	-0.13 _{0.42}	-38.60 _{0.00}	-0.04 _{0.81}	-0.07 _{0.67}	-67.30 _{0.00}
BERTScore-F1 [45]	-0.02 _{0.91}	-0.10 _{0.52}	-31.17 _{0.00}	-0.29 _{0.07}	-0.29 _{0.07}	-50.90 _{0.00}
GPT-4o Aug [16]	0.80 _{0.00}	0.78 _{0.00}	-5.75 _{0.00}	0.61 _{0.00}	0.64 _{0.00}	1.73 _{0.09}
GPT-4o Nov [16]	0.63 _{0.00}	0.61 _{0.00}	-3.45 _{0.00}	0.75 _{0.00}	0.72 _{0.00}	4.76 _{0.00}
o3-mini (high) [33]	0.62 _{0.00}	0.67 _{0.00}	-5.12 _{0.00}	0.47 _{0.03}	0.48 _{0.03}	1.92 _{0.06}
GPT-4o-mini [32]	0.61 _{0.00}	0.55 _{0.01}	-3.34 _{0.00}	0.45 _{0.05}	0.40 _{0.08}	5.27 _{0.00}
DeepSeek-R1 [9]	0.49 _{0.03}	0.43 _{0.06}	-9.70 _{0.00}	0.39 _{0.09}	0.24 _{0.31}	-2.33 _{0.03}
Claude 3.7 Sonnet [1]	0.48 _{0.03}	0.48 _{0.03}	-7.89 _{0.00}	0.66 _{0.00}	0.68 _{0.00}	0.36 _{0.72}
Llama 3.3 70B Instruct [12]	0.68 _{0.00}	0.59 _{0.01}	-1.64 _{0.11}	0.67 _{0.00}	0.58 _{0.01}	8.00 _{0.00}
Qwen 2.5 72B Instruct [35]	0.51 _{0.02}	0.46 _{0.04}	-3.85 _{0.00}	0.54 _{0.01}	0.52 _{0.02}	2.10 _{0.04}
Gemini 2.0 Flash [11]	0.07 _{0.75}	0.10 _{0.68}	0.18 _{0.85}	0.25 _{0.28}	0.21 _{0.39}	10.37 _{0.00}

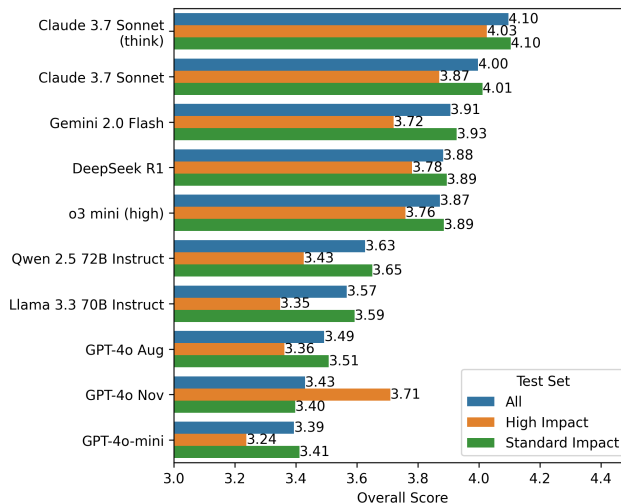
Table 5: Agreement between the expert consensus and GPT-4o Aug for each criterion. “High ($n=3$)” denotes a High-Confidence Subgroup and “Full ($n=8$)” denotes Full Expert Panels. Subscripts indicate p -values.

Criteria	High ($n=3$)		Full ($n=8$)	
	Pearson	Spearman	Pearson	Spearman
Material Appropriateness	0.59 _{0.01}	0.59 _{0.01}	0.44 _{0.05}	0.41 _{0.07}
Equipment Appropriateness	-0.25 _{0.29}	-0.25 _{0.28}	0.10 _{0.68}	0.13 _{0.58}
Procedure Completeness	0.05 _{0.83}	0.09 _{0.71}	0.23 _{0.33}	0.20 _{0.39}
Procedure Similarity	0.41 _{0.07}	0.40 _{0.08}	0.56 _{0.01}	0.50 _{0.02}
Procedure Feasibility	-0.04 _{0.86}	-0.04 _{0.86}	-0.04 _{0.86}	-0.04 _{0.86}
Character. Appropriateness	0.43 _{0.06}	0.42 _{0.07}	0.43 _{0.06}	0.42 _{0.07}
Character. Similarity	0.45 _{0.05}	0.47 _{0.04}	0.09 _{0.72}	0.16 _{0.50}

struggle most with selecting appropriate materials and closely matching the actual synthesis procedures for materials of high-impact papers. The overall score decreased by -0.162 (-4.34%), demonstrating a clear increase in task difficulty for high-impact targets. The experimental results provide valuable insights into the challenges and opportunities in materials synthesis prediction, structured around the following research questions:

Table 6: Mean (σ) absolute and percentage reductions in each evaluation criterion for high-impact materials relative to standard-impact materials

Criteria	Absolute Mean σ	Percentage Mean σ
Materials Appropriateness	-0.310 _{0.092}	-7.98% _{2.500}
Equipment Appropriateness	-0.051 _{0.068}	-1.26% _{1.660}
Procedure Completeness	-0.191 _{0.099}	-5.08% _{2.690}
Procedure Similarity	-0.209 _{0.040}	-7.53% _{1.540}
Procedure Feasibility	-0.037 _{0.073}	-0.93% _{1.750}
Characterization Appropriateness	-0.159 _{0.088}	-3.63% _{2.030}
Characterization Similarity	-0.119 _{0.090}	-3.51% _{2.700}
Overall Score	-0.162 _{0.055}	-4.34% _{1.610}

**Figure 3: Bar chart comparing the overall scores of different LLMs on the full, high-impact, and standard-impact test sets.**

6.3 RQ2: Why Does LLM Performance Differ by Synthesis Process?

Figure 4 demonstrates the mean LLM-as-a-Judge scores for seven evaluation criteria across eleven synthesis processes. Three consistent trends emerge: **1) Organic/biological syntheses are the easiest domain.** The biological column attains the highest overall score (3.89) and leads six of the seven criteria. Organic reactions are typically expressed as a *sequence of well-established named transformations* and can be encoded compactly in line notations such as SMILES; this regularity gives LLMs abundant, structured training signals [2, 8]. **2) Electrochemical syntheses remain challenging.** Electrochemical workflows show the lowest overall score (3.67) and the weakest *Procedure-Similarity* (2.66). Inorganic redox reactions depend on lattice reorganizations, polymorphism, and multi-step

nucleation/growth that are hard to verbalise and poorly covered in literature corpora [5, 6]. The result highlights a critical gap for autonomous inorganic synthesis design. **3) LLMs reproduce what to measure, not how to do it.** *Characterization Appropriateness* is uniformly high (4.35–4.52), whereas *Procedure Similarity* is below 2.9 for every process. Standardized analytical techniques (e.g., XRD, SEM) are easily memorized, yet the procedural ordering of steps is still an open research problem. These findings imply that process-aware fine-tuning or prompt engineering may be necessary: models trained on template-rich organic corpora approach the performance ceiling, whereas electrochemical and mechanochemical domains will benefit from symbolic planners or physics-informed constraints that inject lattice chemistry priors.

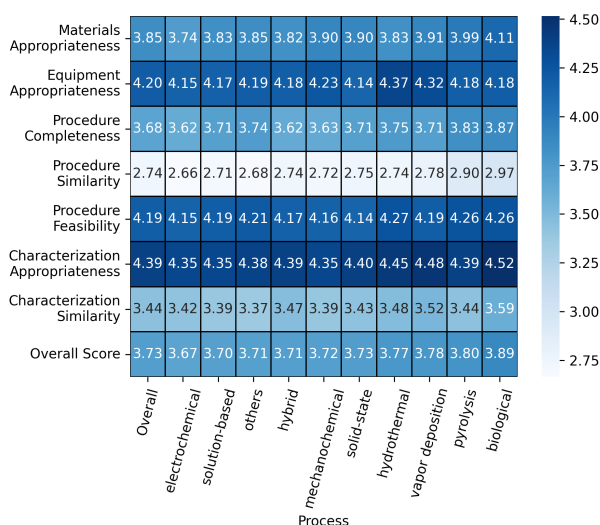


Figure 4: Heatmap illustrating LLM performance across synthesis processes and evaluation criteria.

6.4 RQ3: Does Element-Pair Frequency Affect Performance?

Figure 5 correlates the corpus frequency of target-material element tuples with their overall prediction score. **1) The correlation is positive but weak ($R=0.05$),** indicating that data abundance alone is insufficient. **2) High-frequency binary oxides (Zn–O, Ti–O, Fe–O) achieve ≥ 3.78 despite their chemical diversity,** reflecting the vast and homogeneous oxide literature [39]. Conversely, low-frequency but well-standardized systems such as Ti–N and W–O are also accurately predicted, underscoring the importance of well-defined reaction protocols. **3) Outliers with poor accuracy, e.g., In–Li–Cl or C–H–O,** involve multi-component or hybrid organic/inorganic chemistries whose reaction spaces suffer from the combinatorial explosion and sparse documentation [43]. These findings imply that targeted augmentation of under-represented yet industrially relevant chemistries (chlorides, phosphides, multi-anion systems) is expected to yield the most substantial marginal

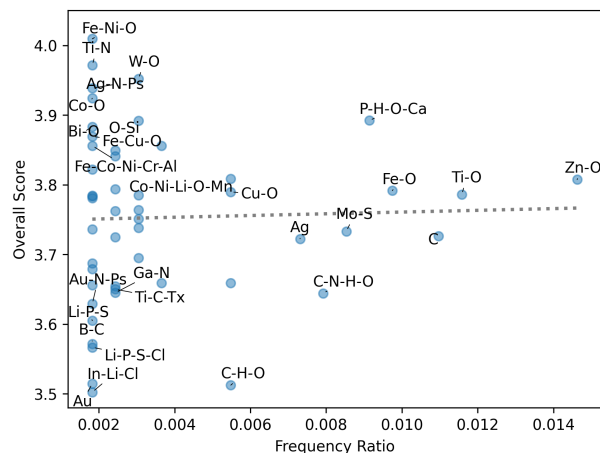


Figure 5: Scatter plot showing the relationship between element combination frequency and overall prediction score.

gain. Moreover, transferring procedure templates learned from data-rich oxides to chemically similar but data-poor chalcogenides may offer a data-efficient route to close the performance gap.

7 FINE-TUNING AND RAG

In this section, we validate the quality of **OMG** training set by measuring its impact on recipe prediction through fine-tuning and retrieval-augmented generation (RAG) [22].

7.1 Experiment Setup

Fine-tuning. For the fine-tuning experiment, we adapted the Qwen 2.5-7B base [35] model, which was selected as a representative smaller model due to the substantial computational cost of fine-tuning larger-scale models via supervised fine-tuning on the **OMG** dataset.

We selected an AdamW [24] optimizer with best-performing hyperparameters, a learning rate of $2e-5$, and trained with a batch size of 128 for three epochs. We used the Hugging Face Transformers [41] library, supplemented by the TRL [15] and DeepSpeed Zero-2 [36] for memory-efficient distributed training. Fine-tuning was executed on four NVIDIA A100 GPUs, employing BF16 mixed-precision [26] arithmetic to maximize throughput while preserving training stability.

Retrieval-Augmented Generation (RAG). To evaluate the impact of retrieval on recipe generation, we implemented an RAG pipeline using OpenAI’s text-embedding-3-large model [31]. For each input X , we retrieved the top- K most similar recipes from the train set based on cosine similarity and included them as references in LLM prompts. We evaluated $K = \{0, 1, 5, 10, 25\}$ to assess the effect of contextual information. RAG experiment ensures a thorough evaluation of both baseline performance and improvements achieved through retrieval augmentation. Due to computational constraints, RAG experiments were conducted on six representative models-Claude 3.7 Sonnet, Claude 3.7 Sonnet thinking, DeepSeek R1, GPT-4o Nov, Gemini 2.0 Flash 001, and o3-mini (high), using only the *High Impact set*.

7.2 RQ4: Does Fine-Tuning on OMG Improve Prediction Performance?

Table 7: Fine-tuning results comparing Qwen-2.5-7B models. Subscripts denote the standard deviation.

Model	High Impact		Standard Impact	
	Mean σ	Max	Mean σ	Max
Qwen2.5-7B-Instruct	2.890 _{0.491}	4.50	3.115 _{0.472}	4.64
Qwen2.5-7B + OMG (ours)	2.914 _{0.678}	4.29	3.175 _{0.568}	4.71

Our fine-tuned model (using only 17K **OMG** examples) outperforms the Qwen 2.5-7B Instruct version, which was trained on over 1 million SFT instances and 100K+ RL examples [35]. This strongly validates our dataset’s high information density and quality - even a tiny domain-specific dataset can effectively guide model learning when the examples are carefully curated. Table 7 presents mean, standard deviation, and maximum scores for high-impact and standard-impact subsets. Fine-tuning yields a modest mean gain of +0.024 (2.890 to 2.914) on high-impact targets, accompanied by an increase in variance (0.491 to 0.678), suggesting greater sensitivity to outlier or complex syntheses. On standard-impact targets, the mean improves by +0.060 (3.115 to 3.175) and the maximum score rises from 4.64 to 4.71, indicating that **OMG** most effectively augments common synthesis recipes. These results imply that domain-specific fine-tuning strengthens model coverage for well-represented materials but may overemphasize frequent patterns. This underscores the need for additional high-impact examples or curriculum scheduling to stabilize performance on rare chemistries.

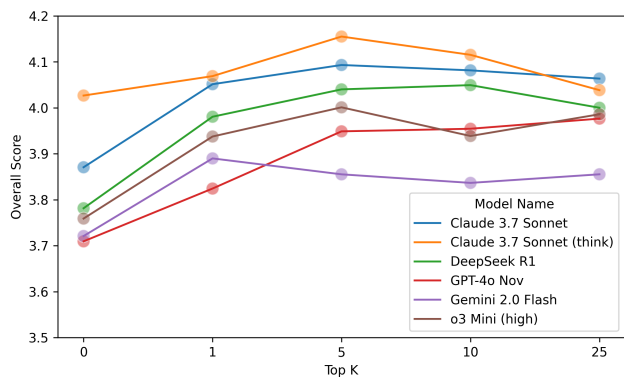


Figure 6: Effect of number of retrieval (K) on overall recipe score for six representative LLMs evaluated on the *High Impact set*.

7.3 RQ5: How Much Contextual Retrieval is Needed to Maximize Recipe Quality?

Figure 6 shows that enriching the prompt with retrieved context steadily raises the overall score from $k=0$ to $k=5$, after which the curve levels off and even dips, implying that a small amount of high-quality evidence is helpful whereas larger bundles mainly

Table 8: Mean differences in criterion-level scores between RAG with $k = 5$ and no retrieval ($k = 0$)

Criteria	Mean σ
Materials Appropriateness	0.245 _{0.569}
Equipment Appropriateness	0.107 _{0.526}
Procedure Completeness	0.307 _{0.692}
Procedure Similarity	0.226 _{0.540}
Procedure Feasibility	0.149 _{0.508}
Characterization Appropriateness	0.205 _{0.564}
Characterization Similarity	0.185 _{0.588}
Overall	0.204 _{0.414}

inject noise. The criterion-wise breakdown in Table 8 confirms that retrieval repairs the parts of the recipe most likely to be underspecified in the base generation: procedure completeness gains +0.307 on average and procedural similarity improves by +0.226, while materials selection also benefits; equipment choice barely moves, indicating that standard instrumentation was already encoded in the model’s prior. Large standard deviations, especially for completeness and materials, underline that some queries receive highly relevant snippets whereas others inherit misleading text.

8 CONCLUSION

We present **OMG**, the large-scale open-access synthesis corpus to date, and **AlchemyBench**, the first benchmark that scores every stage of a materials recipe. By validating an *LLM-as-a-Judge* scheme, we reduce evaluation cost by two orders of magnitude while maintaining expert-level reliability. Experiments reveal: 1) Reasoning-focused LLMs can already draft coherent procedures for well-documented oxide and inorganic targets. 2) Performance collapses for electrochemical or low-frequency element pairs, indicating an urgent need for process-aware fine-tuning and data augmentation. 3) Retrieval-augmented prompts supply the missing procedural details, but only when context is filtered for quality. These findings underscore the importance of combining advanced reasoning architectures with adaptive retrieval strategies for materials science tasks, laying the foundation for interdisciplinary innovation and accelerating progress in data-driven and fully-automated materials discovery.

Acknowledgements

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. II220680, Abductive inference framework using omni-data for understanding complex causal relations, 20%), (RS-2025-02263277, Development of AGI platform for multisensory social robot, 20%), (IITP-2025-RS-2023-00255968, 20%) and the National R&D Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science and ICT (No. RS-2024-00407282, 20%), and (No. RS-2024-00444182, 20%).

GENAI USAGE DISCLOSURE

Following the CIKM 2025 submission guidelines, we disclose our use of generative AI tools in preparing this manuscript.

- **Microsoft Copilot** (Copilot) was employed exclusively as a *coding assistant* to accelerate the development of data-processing scripts and algorithmic prototypes. All code suggestions from Copilot were manually reviewed, adapted, and tested by the authors; no AI-generated code was used without human verification.
- **Grammarly** (Grammarly) and **Writefull** (Writefull) were used as writing assistants throughout drafting, including grammar and spelling corrections, sentence rephrasing for clarity and conciseness, and style-guide conformance. All suggested edits from these tools were carefully reviewed and, where appropriate, manually revised to preserve authorial voice and ensure technical accuracy. AI generated no substantive conceptual content or experimental analysis.
- **Scope of AI-Generated Content:** Any text segments directly influenced by AI recommendations are marked in draft comments and have been replaced or refined by the authors. The final manuscript contains only author-verified prose.

References

- [1] Anthropic. 2025. Claude 3.7 Sonnet System Card. <https://assets.anthropic.com/m/785e231869ea8b3b/original/claude-3-7-sonnet-system-card.pdf>.
- [2] Evan R Antoniuk, Gowoon Cheon, George Wang, Daniel Bernstein, William Cai, and Evan J Reed. 2023. Predicting the synthesizability of crystalline inorganic materials from the data of known material compositions. *npj Computational Materials* 9, 1 (2023), 155.
- [3] Artifex Software. 2024. *PyMuPDF4LLM: PDF Text Extraction Library for LLM Applications*. Accessed: January 2024.
- [4] Authors Alliance. 2024. *Text and Data Mining Under U.S. Copyright Law: Landscape, Flaws & Recommendations*. Technical Report. <https://www.authorsalliance.org/wp-content/uploads/2024/11/Text-and-Data-Mining-Report-102024.pdf>
- [5] Murathan Aykol, Joseph H Montoya, and Jens Hummelshøj. 2021. Rational solid-state synthesis routes for inorganic materials. *Journal of the American Chemical Society* 143, 24 (2021), 9244–9259.
- [6] Juan R Chamorro and Tyrel M McQueen. 2018. Progress toward solid state synthesis by design. *Accounts of chemical research* 51, 11 (2018), 2918–2925.
- [7] Yuan Chiang, Elvis Hsieh, Chia-Hong Chou, and Janosh Riebesell. 2024. LLaMP: Large Language Model Made Powerful for High-fidelity Materials Knowledge Retrieval and Distillation. arXiv:2401.17244 [cs.CL] <https://arxiv.org/abs/2401.17244>
- [8] Elias J Corey, Richard D Cramer III, and W Jeffrey Howe. 1972. Computer-assisted synthetic analysis for complex molecules. Methods and procedures for machine generation of synthetic intermediates. *Journal of the American Chemical Society* 94, 2 (1972), 440–459.
- [9] DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jia Shi Li, Jiawei Wang, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Meng Li, Miaojuan Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiusi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shutong Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Shengfeng Ye, Tao Yun, Tian Pei, Tianyu Sun, T. Wang, Wangding Zeng, Wanji Zhao, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanhong Xu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yuxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. arXiv:2501.12948 [cs.CL] <https://arxiv.org/abs/2501.12948>
- [10] Alexander Dunn, Qi Wang, Alex Ganose, Daniel Dopp, and Anubhav Jain. 2020. Benchmarking materials property prediction methods: the Matbench test set and Automatminer reference algorithm. *npj Computational Materials* 6, 1 (2020), 138.
- [11] Google. 2025. Gemini 2.0 Flash | Generative AI on Vertex AI | Google Cloud — cloud.google.com. <https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-0-flash>.
- [12] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).
- [13] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2025. A Survey on LLM-as-a-Judge. arXiv:2411.15594 [cs.CL] <https://arxiv.org/abs/2411.15594>
- [14] Guannan Huang, Yani Guo, Ye Chen, and Zhengwei Nie. 2023. Application of machine learning in material synthesis and property prediction. *Materials* 16, 17 (2023), 5977.
- [15] Huggingface. [n. d.]. GitHub - huggingface/trl: Train transformer language models with reinforcement learning. — github.com. <https://github.com/huggingface/trl>.
- [16] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).
- [17] Shuyi Jia, Chao Zhang, and Victor Fung. 2024. LLMatDesign: Autonomous Materials Discovery with Large Language Models. *arXiv preprint arXiv:2406.13163* (2024).
- [18] Seongmin Kim, Yousung Jung, and Joshua Schrier. 2024. Large language models for inorganic synthesis predictions. *Journal of the American Chemical Society* 146, 29 (2024), 19654–19659.
- [19] Olga Kononova, Haoyan Huo, Tanjin He, Ziqin Rong, Tiago Botari, Wenhao Sun, Vahe Tshitoyan, and Gerbrand Ceder. 2019. Text-mined dataset of inorganic materials synthesis recipes. *Scientific data* 6, 1 (2019), 203.
- [20] Shrinidhi Kumbhar, Venkatesh Mishra, Kevin Coutinho, Divij Handa, Ashif Iqbal, and Chitta Baral. 2025. Hypothesis Generation for Materials Discovery and Design Using Goal-Driven and Constraint-Guided LLM Agents. arXiv:2501.13299 [cs.CL] <https://arxiv.org/abs/2501.13299>
- [21] Ge Lei, Ronan Docherty, and Samuel J Cooper. 2024. Materials science in the era of large language models: a perspective. *Digital Discovery* (2024).
- [22] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.
- [23] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*. 74–81.
- [24] Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. arXiv:1711.05101 [cs.LG] <https://arxiv.org/abs/1711.05101>
- [25] Amil Merchant, Simon Batzner, Samuel S Schoenholz, Murathan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. 2023. Scaling deep learning for materials discovery. *Nature* 624, 7990 (2023), 80–85.
- [26] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. 2017. Mixed precision training. *arXiv preprint arXiv:1710.03740* (2017).
- [27] Vaibhav Mishra, Somaditya Singh, Mohd Zaki, Hargun Singh Grover, Santiago Miret, NM Anoop Krishnan, et al. [n. d.]. LLaMat: Large Language Models for Materials Science. In *AI for Accelerated Materials Design-Vienna 2024*.
- [28] Gyoung S Na. 2023. Artificial intelligence for learning material synthesis processes of thermoelectric materials. *Chemistry of Materials* 35, 19 (2023), 8272–8280.
- [29] Tim O’Gorman, Zach Jensen, Sheshera Mysore, Kevin Huang, Rubayyat Mahbub, Elsa Olivetti, and Andrew McCallum. 2021. MS-Mentions: Consistently Annotating Entity Mentions in Materials Science Procedural Text. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Online and Punta Cana, Dominican

- Republic, 1337–1352. doi:10.18653/v1/2021.emnlp-main.101
- [30] Elsa A Olivetti, Jacqueline M Cole, Edward Kim, Olga Kononova, Gerbrand Ceder, Thomas Yong-Jin Han, and Anna M Hiszpanski. 2020. Data-driven materials research enabled by natural language processing and information extraction. *Applied Physics Reviews* 7, 4 (2020).
 - [31] OpenAI. 2022. Introducing text and code embeddings. <https://openai.com/index/introducing-text-and-code-embeddings/>. [Accessed 11-02-2025].
 - [32] OpenAI. 2024. GPT-4o mini: advancing cost-efficient intelligence — openai.com. <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>.
 - [33] OpenAI. 2025. OpenAI o3-mini — openai.com. <https://openai.com/index/openai-o3-mini/>.
 - [34] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.
 - [35] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 Technical Report. arXiv:2412.15115 [cs.CL] <https://arxiv.org/abs/2412.15115>
 - [36] Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20: International Conference for High Performance Computing, Networking, Storage and Analysis*. IEEE, 1–16.
 - [37] Patrick E Shrout and Joseph L Fleiss. 1979. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin* 86, 2 (1979), 420.
 - [38] Yu Song, Santiago Miret, and Bang Liu. 2023. Matsci-nlp: Evaluating scientific language models on materials science language tasks using text-to-schema modeling. *arXiv preprint arXiv:2305.08264* (2023).
 - [39] Wenhao Sun and Nicholas David. 2025. A critical reflection on attempts to machine-learn materials synthesis insights from text-mined literature recipes. *Faraday Discussions* (2025).
 - [40] Zheren Wang, Olga Kononova, Kevin Cruse, Tanjin He, Haoyan Huo, Yuxing Fei, Yan Zeng, Yingzhi Sun, Zijian Cai, Wenhao Sun, et al. 2022. Dataset of solution-based inorganic materials synthesis procedures extracted from the scientific literature. *Scientific data* 9, 1 (2022), 231.
 - [41] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* (2019).
 - [42] Pengcheng Xu, Xiaobo Ji, Minjie Li, and Wencong Lu. 2023. Small data machine learning in materials science. *npj Computational Materials* 9, 1 (March 2023). doi:10.1038/s41524-023-01000-z
 - [43] Pengcheng Xu, Xiaobo Ji, Minjie Li, and Wencong Lu. 2023. Small data machine learning in materials science. *npj Computational Materials* 9, 1 (2023), 42.
 - [44] Angel Yanguas-Gil, Matthew T Dearing, Jeffrey W Elam, Jessica C Jones, Sungjoon Kim, Adnan Mohammad, Chi Thang Nguyen, and Bratin Sengupta. 2024. Benchmarking large language models for materials synthesis: the case of atomic layer deposition. *arXiv preprint arXiv:2412.10477* (2024).
 - [45] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675* (2019).
 - [46] Yu Zhao, Dongru Gao, Ruxin Guan, Hongwei Li, Ning Li, Guixian Li, and Shiyu Li. 2020. Synthesis of a three-dimensional cross-linked Ni–V 2 O 5 nanomaterial in an ionic liquid for lithium-ion batteries. *RSC advances* 10, 64 (2020), 39137–39145.
 - [47] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* 36 (2023), 46595–46623.