A Supplementary Information

A.1 Data

This section outlines details for the dataset (Appendix A.1.1) as well as the data curation (Appendix A.1.2) presented in this work.

A.1.1 Dataset statistics

We classify each material into a set of predetermined material categories and synthesis methods, as determined by the recommendation of domain experts.

Material categories With the goal of covering practically the entire space of material science synthesis, the following material categories were chosen by domain experts of our group and are employed in this work: metals & alloys, ceramics & glasses, polymers & soft matter, composites, semiconductors & electronic, nanomaterials, two-dimensional materials, framework & porous materials, biomaterials & biological, liquid materials, hybrid & organic-inorganic, functional materials & catalysts, energy & sustainability, smart & responsive materials, emerging & quantum materials. Any category not covered in the list is assigned the label "other".

Synthesis methods Similarly, the following material categories were chosen by domain experts of our group and are employed in this work: PVD, CVD, arc discharge, ball milling, spray pyrolysis, electrospinning, sol-gel, hydrothermal, solvothermal, precipitation, coprecipitation, combustion, microwave-assisted, sonochemical, template-directed, solid-state, flux growth, float zone & Bridgman, arc melting & induction melting, spark plasma sintering, electrochemical deposition, chemical bath deposition, liquid-phase epitaxy, self-assembly, atomic layer deposition, molecular beam epitaxy, pulsed laser deposition, ion implantation, lithographic patterning, wet impregnation, incipient wetness impregnation, mechanical mixing, solution-based, mechanochemical. Any category not covered in the list is assigned the label "other".

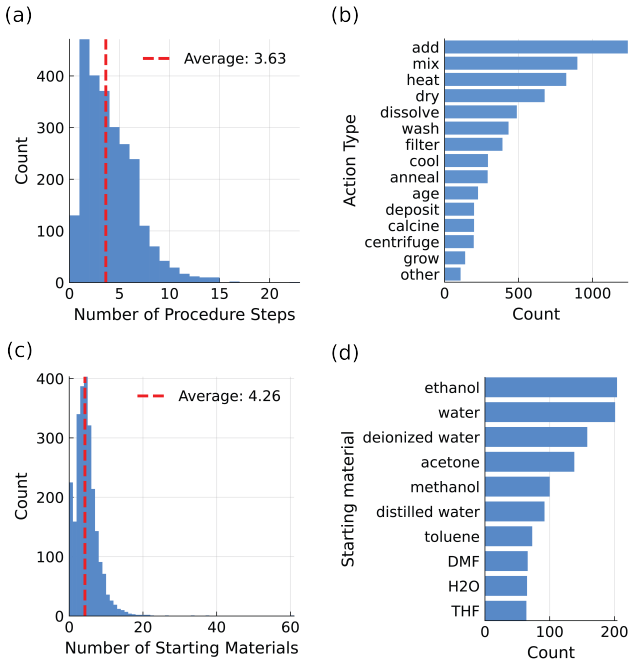

Figure 4: Statistics of the dataset evaluated in this work. **(a)** Distribution of action steps and **(b)** the 15 most common actions. **(c)** Distribution of the number of starting materials and **(d)** the 10 most common starting materials. Note that similarly to material identifiers, starting materials are not standardized.

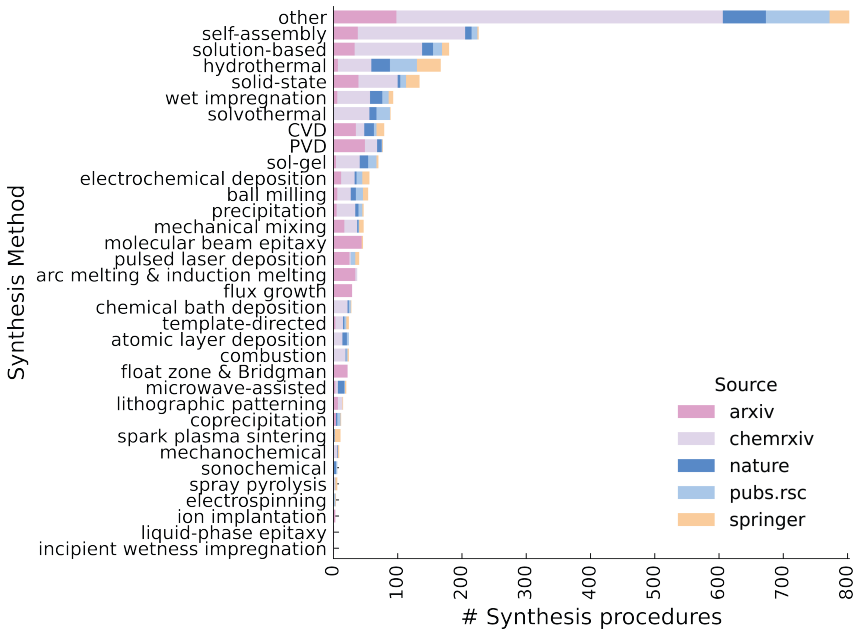

Figure 5: Synthesis procedures and methods for the evaluation set, colored according to the source of the underlying publication (arXiV, ChemRxiv, OMG24).

Note that due to the costs of creating the whole dataset which is expected to contain 100-150k synthesis procedures, we perform all evaluations on a random subset of 2.5k synthesis procedures (526 stemming from the arXiV, 1252 ChemRxiv, 706 omg24 (239 Nature, 279 RSC, 188 Springer). While this split is not stratified with respect to the entire corpus, we claim that it is a representative sample (approx. 2-2.5%) that covers a broad array of synthesis methods, see Table 5 and Table 6. We are currently rolling out the inference pipeline to the whole corpus of 81k publications.

A.1.2 Data acquisition

arXiV From over two million articles on arXiV in total, we fetched 381116 publications in the category `cond-mat` from 1992 to April 2025. We filtered down the corpus to 62,267 publications that contain synthesis procedures by parsing the PDF with Marker and calling `Mistral-Small-3.1-24B-Instruct-2503` on a cluster of 8xA100-PG509-200 with 40GB of memory each. The text from the PDF (if length larger the max tokens, chunk paper) is passed to the LLM to return whether it contains a synthesis procedure, the material name and category, see Appendix A.4.

ChemRxiv From over 30000 articles with the cutoff date of June 2025, we fetched 2910 publications in the categories Solid State Chemistry, Solution Chemistry, Solvates, Spectroscopy (Inorg.), Structure, Supramolecular Chemistry (Inorg.), Supramolecular Chemistry (Org.), Surface, Surfactants, Thermal Conductors and Insulators, Thin Films, Wastes, Water Purification, with the ChemRxiv API. We obtain 1500 papers with synthesis procedures. If available, a supplementary file is appended to the main text.

Open Materials Guide 2024 (OMG24) The data collection and curation from the Semantic Scholar API is described in [21]. It contains 17667 synthesis procedures with ten different synthesis types from open access publications. We fetched the PDFs from the URLS provided in the published dataset, downloaded it and proceeded with parsing the text and images. As the papers in OMG24 are already pre-filtered to contain synthesis procedures, no filtering step is needed.

PDF post-processing To extract text and figures from PDFs obtained from the arXiV, we use `marker-pdf`, an open-source library, with `Gemini 2.0 flash` (`gemini-2.0-flash`). We strip the images from the text, which is converted into Markdown format, and save the images separately,

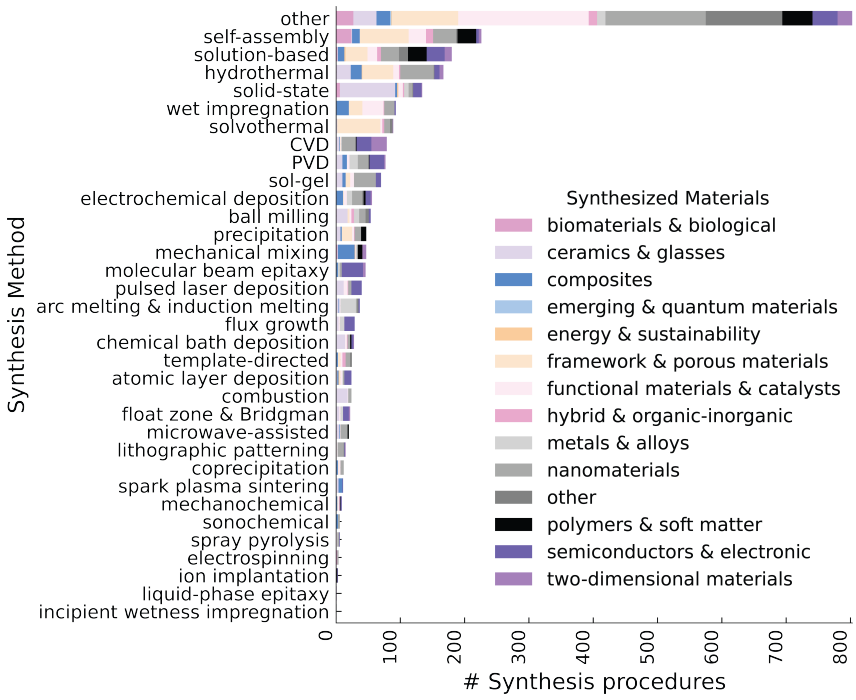

Figure 6: Synthesis procedures and methods for the evaluation set, colored according to the material category.

but such that they can be reinserted into the Markdown text. For the ChemRxiv and OMG24, we used `Mistral-OCR` (`mistral-ocr-latest`) to extract images and text in Markdown format. We empirically tested `Docling` [36], an open source alternative to Mistral-OCR, and found Mistral-OCR to empirically perform better and infer results faster. For post-processing the text, we removed markdown image identifiers and the References section (= 50 lines after the heading References with regex).

Conversely, entries for which no valid synthesized material was found (23%), the name consisted of a character and/or symbol only (12%) or the material was described with an unclear identifier ("Intermediate 1", "8a", "Compound B" etc.) (0.3%) were subsequently filtered out to maintain data quality. This high dropout rate highlights the need to standardize material identifiers to further make the database properly searchable and interoperable. Lastly, entries where the extraction failed according to the LLM-as-a-judge (*vide infra*, a materials extraction score equal to one) were filtered out (13%), likely due to the complex ontology enforced.

A.2 Synthesis Extraction

Manual annotations Seven material scientists cross-manually annotated a total of 35 papers ([37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 60, 61, 62, 63, 64, 65, 66, 67] by inferring synthesis procedures from a sample picked at random among each of the following sources: arXiV, ChemRxiv, OMG24 (1 to 1 ratio, stratified sampling). The synthesis procedures were manually reviewed for correctness, completeness, and adherence to a pre-defined structured ontology. Note that this process ensured the relevant information was extracted as it was in the text, and didn't aim to directly assess scientific accuracy. To the material scientists' capacity, where relevant but ambiguous terms from the experimental workflows needed to be assessed, more than one annotator was consulted and a consensus was reached in order to maintain the consistency throughout the process.

Each validation assessed whether the LLM-extracted synthesis procedures were consistent with the original text. The annotators noted down any missing, incorrect or hallucinated content generated and attributed detailed scores for each procedure. A total of seven scoring criteria were used, ranging from 1 (poor) to 5 (excellent) in 0.5 increments:

- **Structural completeness score**: Coverage of ontology-relevant information, including materials, synthesis steps, equipment, conditions, etc.

- **Material extraction score**: accuracy and completeness of the extracted materials, including names, quantities, units, and purities.

- **Process steps score**: correctness and organization of the procedural steps, including the sequence and classification of synthesis actions.

- **Equipment extraction score**: completeness and accuracy in identifying experimental apparatus, including vendor names and operational settings where available.

- **Conditions extraction score**: correctness of temperature, pressure, duration, and atmospheric conditions, along with unit consistency.

- **Semantic accuracy score**: the degree to which the structured extraction preserved the scientific meaning and contextual integrity of the original description.

- **Format compliance score**: adherence of the structured data to the ontology schema and data type requirements.

Finally, an **overall score** was computed as the mean of the individual criteria, with a final reasoning field summarizing strengths, weaknesses, and suggestions for improvement.

A.2.1 Ontology

Figure 7 and Table 1 show the ontology developed in this work. We abstracted a *broad* synthesis procedure as a sequence of steps with actions, conditions, equipment and an associated material, as well as starting materials. Note that in the library released in this work, the ontology can be adapted to custom cases, e.g. specialized syntheses for catalysts or polymers. The ontology can be adapted from the `GeneralSynthesisOntology` class.

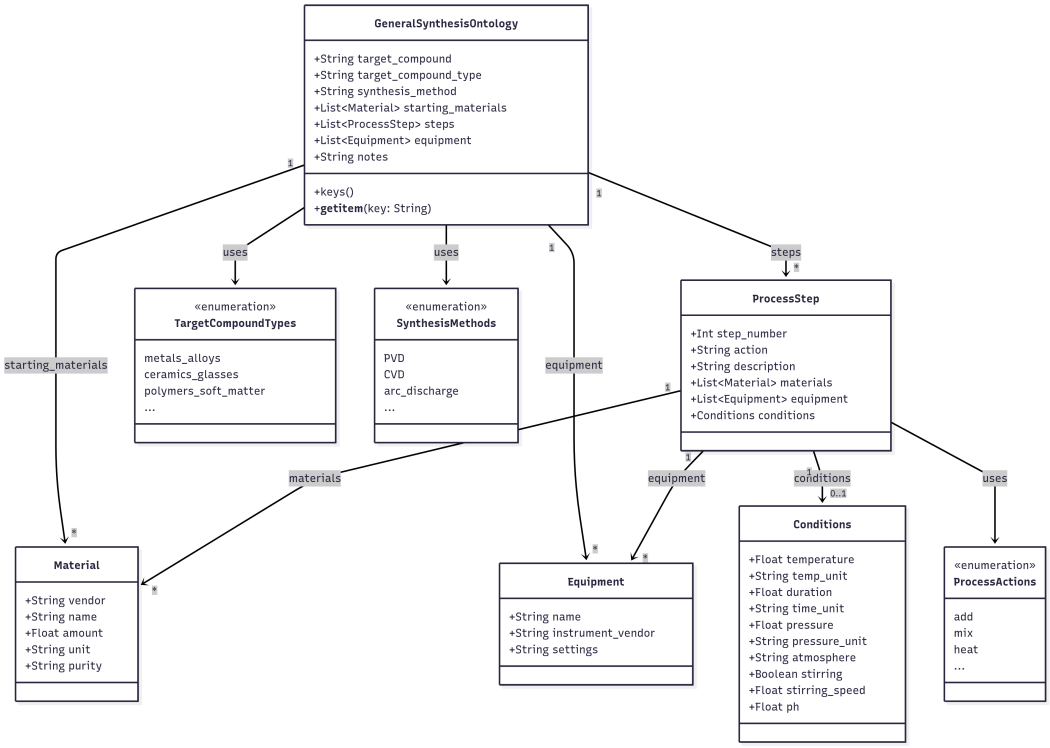

Figure 7: Visual representation of the hierarchical ontology for structuring synthesis procedures. The ontology organizes information from a global level (target compound, synthesis method) down to sequential process steps. Each step encapsulates detailed information about the specific actions, materials, equipment, and conditions involved, ensuring data consistency and machine-readability (Table 1).

A.2.2 Domain expert – LLM as a judge comparison

The high Spearman correlation demonstrates that the LLM has demonstrated the ability to distinguish better from worse extractions, which is practically valuable as the rank-order of scores between humans and LLM-judge will be similar. The exact agreement is lower (Cohen's $\kappa = 0.44$), but this is a result of calibration differences rather than fundamental disagreement. Discrepancies typically arise when literature descriptions are vague or incomplete — experts may infer plausible synthesis details, whereas the LLM more strictly penalizes under-specified inputs.

Example 1: Lower Agreement (Material: Au–OLC) This paper demonstrated significant disagreement between the LLM and human validations, with the LLM consistently overestimating extraction quality. The most substantial disagreements occurred in Structural Completeness and Process Steps (both 2.0 point differences), stemming from fundamental misidentification of key synthesis components. Most critically, the extraction incorrectly labeled the gold precursor as "chloroplatinic acid"—a platinum-containing compound that would be chemically impossible to use for gold nanoparticle synthesis. Additionally, the system missed essential materials including water and mixed acid, and misclassified the annealing and hydrothermal treatment as a generic "heat" action rather than the specific synthesis method. In contrast, the other metal-OLC materials (Pt-OLC, Pd-OLC, Ag-OLC) extracted from the same paper achieved higher overall scores, suggesting that the extraction difficulties were specific to the Au-OLC synthesis description rather than a systematic issue with the paper's clarity. The LLM's overconfidence in its extraction quality, despite these fundamental chemical and procedural errors, highlights the critical importance of human validation for ensuring extraction accuracy in complex nanomaterial synthesis procedures.

Table 1: Detailed structure of the `GeneralSynthesisOntology` scheme for the standardized representation of asynthesis procedure. Note that the type (material category) and synthesis method are chosen from a pre-determined list of verbs. The General Synthesis Ontology contains the target compound, synthesis method, overall materials and equipment. The Process Steps object is sequential and contains ordered operations with specific actions, local materials, equipment, and conditions. Materials (Chemical identity, quantities, specifications, and vendor), Equipment (Instrumentation with settings and vendor information), Conditions (Environmental parameters: temperature, time, pressure, atmosphere, pH) are set.

Component	Attributes	Description & Examples
Target Compound	compound	Chemical composition and description
	type	Material category: metals & alloys, ceramics, nanomaterials, polymers, semiconductors, etc.
	synthesis method	Technique: sol-gel, hydrothermal, CVD, precipitation, electrodeposition, etc.
	notes	Additional observations or variations
Material	name	Chemical name (e.g., Nickel Nitrate, Deionized Water)
	amount	Quantity used (numeric value)
	unit	Mass (g, mg), Volume (mL, L), Molar (mol, mmol), Concentration (M, mM), etc.
	vendor	Supplier information
	purity	Grade specification (99%, ACS grade, etc.)
Equipment	name	Instrument type (autoclave, tube furnace, magnetic stirrer)
	vendor	Manufacturer (Thermo Fisher, Agilent, Bruker, etc.)
	settings	Operating parameters (500 rpm, heating rate 5°C/min)
Conditions	temperature	Process temperature with units (°C, K, °F)
	duration	Time period with units (h, min, s, days)
	pressure	Applied pressure with units (atm, bar, Pa, torr)
	atmosphere	Gas environment (air, N_2, H_2, Ar, vacuum)
	stirring	Boolean and speed (rpm)
	pH	Solution acidity/basicity
Process Step	step number	Sequential order in procedure
	action	Primary operation: add, mix, heat, cool, reflux, age, filter, wash, dry, etc.
	description	Detailed procedure text
	materials	List of materials used in this step
	equipment	List of equipment used in this step
	conditions	Environmental parameters for this step

Example 2: High Agreement (Material: Fluorapatite–Titania Nanocomposite) This example demonstrates excellent agreement between LLM and human evaluations, with perfect consensus across six of seven criteria and only a minor 0.5-point difference in Semantic Accuracy. The extraction successfully captured all key aspects of the mechano-chemical synthesis procedure, correctly identifying the starting materials ($CaHPO_4$, $Ca(OH)_2$, CaF_2, and TiO_2), process steps (mixing, ball milling, annealing), and reaction conditions. The LLM accurately extracted specific parameters such as the 20 wt% TiO_2 content, 600 rpm milling speed, and 700°C annealing temperature, while properly classifying the synthesis method as ball milling followed by thermal treatment.

Furthermore, the LLM only evaluates synthesis procedures that are extracted, and does not point out procedures that failed to extract.

Table 2: Comparing domain expert evaluations to LLM-as-a-judge. μ_{exp}, $\mu_{1/2,exp}$ and σ_{exp} refer to the mean, median and standard deviation for all six annotators and μ_{LLM}, $\mu_{1/2,LLM}$ and σ_{LLM} to the mean, median and standard deviation of the LLM (`Gemini-2.0-flash`), respectively.

Criterion	Spearman	p-value	Cohen	ICC(2,1)	ICC(3,1)	μ_{exp}	$\mu_{1/2,exp}$	σ_{exp}	μ_{LLM}	$\mu_{1/2,LLM}$	σ_{LLM}
Structural Completeness	0.4209	0.0004	0.2029	0.2286	0.2304	4.12	4.00	0.65	4.02	4.00	0.40
Material Extraction	0.7107	0.0002	0.5790	0.5996	0.5964	4.08	4.00	0.89	4.11	4.00	0.59
Process Steps	0.5547	0.0002	0.2867	0.2620	0.2626	4.15	4.25	0.82	4.27	4.25	0.55
Equipment Extraction	0.5842	0.0002	0.6287	0.6229	0.6325	4.05	4.50	1.19	3.80	4.00	1.18
Conditions Extraction	0.6201	0.0002	0.4747	0.4283	0.4565	4.27	4.00	0.70	4.01	4.00	0.68
Semantic Accuracy	0.5407	0.0002	0.3919	0.4170	0.4133	4.39	4.50	0.64	4.39	4.50	0.38
Format Compliance	0.2690	0.0350	0.1129	0.2141	0.2137	4.77	5.00	0.53	4.83	5.00	0.30
Overall	**0.7195**	**0.0002**	**0.4407**	**0.5411**	**0.5399**	**4.25**	**4.30**	**0.52**	**4.20**	**4.25**	**0.42**

Table 3: Evaluation scores for a low-agreement synthesis procedure extraction for Au-OLC from paper id 9a889c1a671fd3cae48285eaa95069d189d02fe3443.

Criterion	Human	LLM	Difference
Structural Completeness	2.0	4.0	2.0
Material Extraction	2.0	3.0	1.0
Process Steps	2.0	4.0	2.0
Equipment Extraction	5.0	4.0	1.0
Conditions Extraction	5.0	4.5	0.5
Semantic Accuracy	2.0	3.5	1.5
Format Compliance	4.0	5.0	1.0
Overall	3.1	4.0	0.9

A.2.3 Scaling LLM-as-a-judge across the dataset

Figure 8, Figure 9, Figure 11, Figure 10, Figure 12, Figure 13, Table 5 and Table 6 show the performance of LLM-as-a-judge across the dataset. For the sample on which we assess human–LLM agreement ($n = 66$), we report Spearman rank correlations (ρ) between human and model scores, but compute their p-values using a permutation test (10,000 resamples, two-sided) rather than relying on the standard asymptotic approximation. This choice is motivated by the modest sample size and the bounded, quasi-ordinal nature of the scores, which induce many ties and can render asymptotic p-values anticonservative and unreliable. As the SciPy documentation recommends,[3] "for small samples, consider performing a permutation test instead of relying on the asymptotic p-value," especially when ties and discrete data violate large-sample assumptions. The permutation procedure generates the exact finite-sample null distribution of ρ by permuting only one input (human scores) relative to the other while preserving marginal distributions. This approach provides valid inference under exchangeability, naturally handles ties, and ensures robust significance testing even with small, discrete datasets.

[3] `https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html`

Table 4: Evaluation scores for a high-agreement synthesis procedure extraction for Fluorapatite–Titania Nanocomposite from paper id ccc7c5d70ae3ca3f9e975d0dc3b4d631586c1586.

Criterion	Human	LLM	Difference
Structural Completeness	4.0	4.0	0.0
Material Extraction	4.0	4.0	0.0
Process Steps	4.5	4.5	0.0
Equipment Extraction	4.0	4.0	0.0
Conditions Extraction	4.5	4.5	0.0
Semantic Accuracy	4.0	4.5	0.5
Format Compliance	5.0	5.0	0.0
Overall	4.4	4.3	0.1

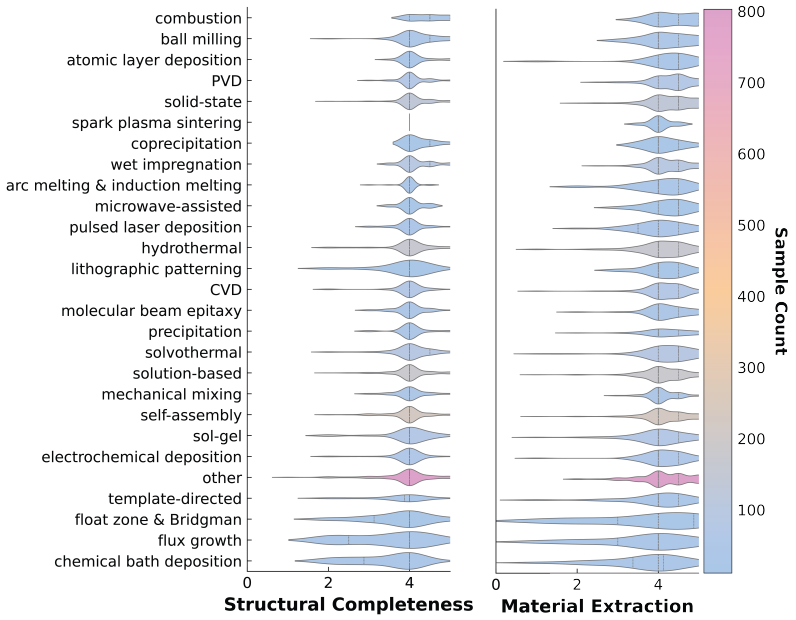

Figure 8: Distribution of LLM-judged overall extraction scores across different synthesis methods (structural completeness and material extraction score). See Table 5 for the full score overview. Each violin plot shows the probability density of the scores for a given synthesis type.

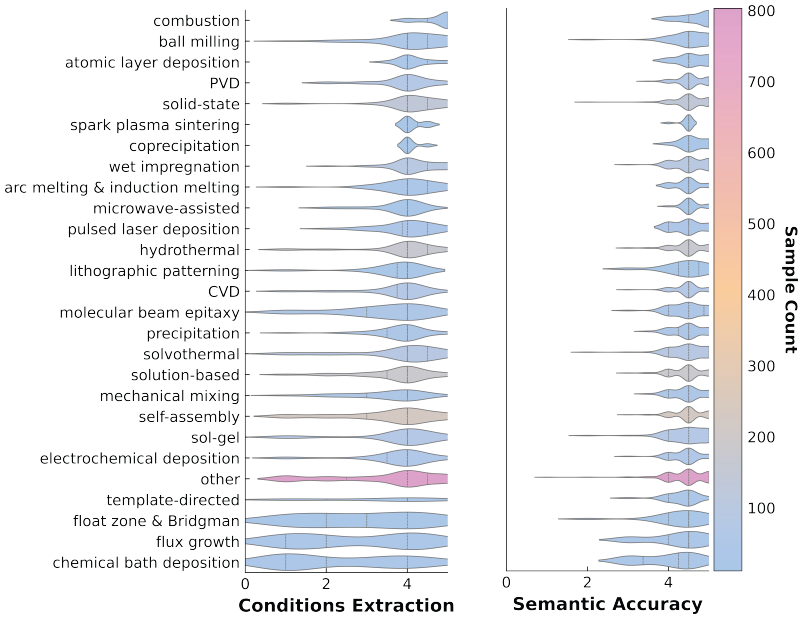

Figure 9: Distribution of LLM-judged overall extraction scores across different synthesis methods (condition extraction and semantic accuracy score). See Table 5 for the full score overview. Each violin plot shows the probability density of the scores for a given synthesis type.

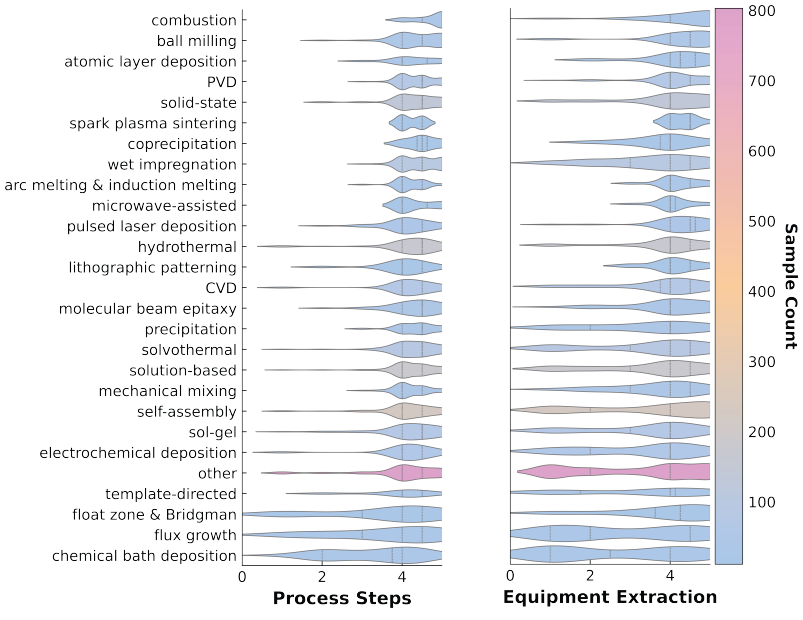

Figure 10: Distribution of LLM-judged overall extraction scores across different synthesis methods (process steps and equipment extraction score). See Table 5 for the full score overview. Each violin plot shows the probability density of the scores for a given synthesis type.

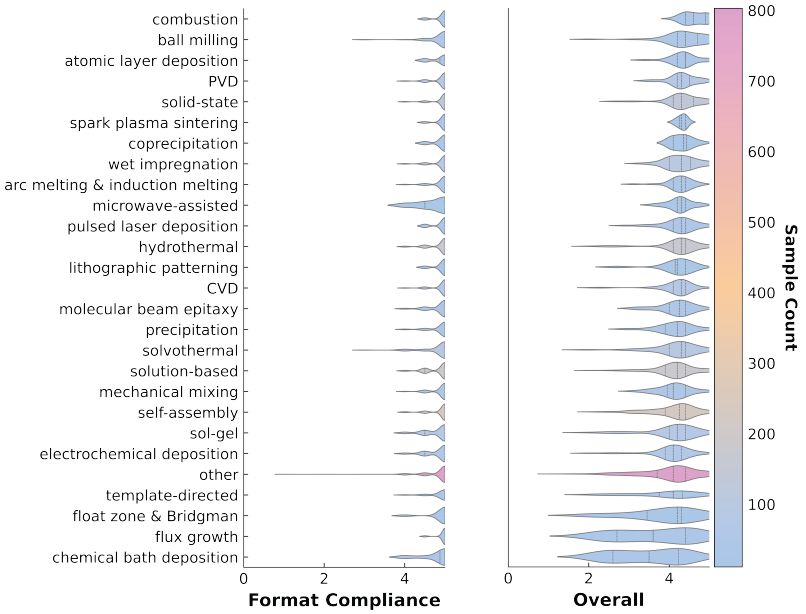

Figure 11: Distribution of LLM-judged overall extraction scores across different synthesis methods (format compliance and overall score). See Table 5 for the full score overview. Each violin plot shows the probability density of the scores for a given synthesis type.

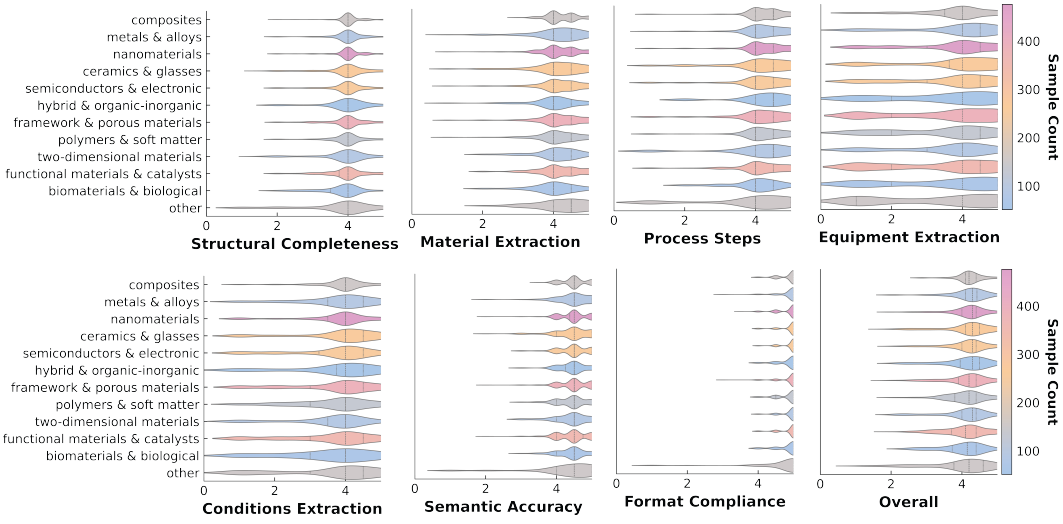

Figure 12: Distribution of LLM-judged overall extraction scores across different material classes. See Table 6 for a complete overview. Each violin plot shows the probability density of the scores for a given material category.

Table 5: Average LLM-judged extraction scores for the most frequent synthesis methods in the evaluated dataset subset ($N = 2483$ procedures). Scores are reported as mean ± standard deviation on a 1–5 scale. The Overall Score is the average of all seven evaluation criteria.

Synthesis method	Structural completeness	Material completeness	Process steps	Equipment extraction	Condition extraction	Semantic accuracy	Format compliance	Overall score	Count
other	3.85±0.73	4.14±0.65	4.00±0.99	3.22±1.55	3.47±1.31	4.42±0.57	4.83±0.43	3.99±0.70	803
self-assembly	3.94±0.50	4.16±0.58	4.13±0.75	3.56±1.53	3.56±1.16	4.49±0.39	4.89±0.26	4.10±0.56	226
solution-based	4.01±0.48	4.12±0.57	4.23±0.61	3.50±1.32	3.71±0.90	4.42±0.40	4.84±0.28	4.12±0.51	180
hydrothermal	3.99±0.57	4.09±0.70	4.17±0.86	3.88±1.06	3.89±0.93	4.47±0.41	4.87±0.27	4.20±0.59	167
solid-state	4.09±0.42	4.29±0.56	4.29±0.61	3.96±1.10	4.13±0.76	4.54±0.41	4.92±0.22	4.32±0.44	134
wet impregnation	4.15±0.37	4.23±0.47	4.42±0.46	3.49±1.20	4.17±0.60	4.53±0.40	4.91±0.23	4.28±0.38	92
solvothermal	4.03±0.52	4.21±0.69	4.26±0.62	3.47±1.42	3.80±1.11	4.47±0.49	4.84±0.37	4.15±0.57	89
CVD	3.96±0.45	4.16±0.55	4.18±0.74	3.79±1.11	3.71±0.87	4.47±0.34	4.92±0.22	4.18±0.46	79
PVD	4.06±0.34	4.32±0.49	4.34±0.42	4.14±0.78	3.92±0.71	4.57±0.34	4.90±0.23	4.32±0.33	77
sol-gel	3.94±0.65	4.00±0.71	4.20±0.77	3.43±1.30	3.76±1.09	4.46±0.53	4.81±0.31	4.09±0.64	70
electrochemical deposition	3.91±0.49	4.14±0.59	4.12±0.82	3.26±1.36	3.74±0.92	4.41±0.37	4.84±0.29	4.06±0.51	56
ball milling	4.17±0.50	4.21±0.56	4.38±0.60	4.36±0.93	4.07±0.87	4.51±0.52	4.88±0.34	4.37±0.52	54
precipitation	4.03±0.38	4.20±0.58	4.35±0.47	3.36±1.35	3.82±0.68	4.48±0.38	4.89±0.25	4.16±0.44	47
mechanical mixing	4.01±0.38	4.11±0.36	4.21±0.41	3.67±1.07	3.46±0.93	4.46±0.37	4.90±0.22	4.12±0.39	47
molecular beam epitaxy	4.00±0.37	4.21±0.54	4.30±0.63	3.96±1.01	3.38±1.10	4.46±0.43	4.87±0.27	4.17±0.42	46
pulsed laser deposition	3.99±0.35	4.01±0.63	4.11±0.61	4.26±0.78	3.99±0.67	4.46±0.36	4.91±0.19	4.25±0.41	40
arc & induction melting	3.99±0.22	4.07±0.69	4.20±0.36	4.22±0.49	4.01±0.75	4.47±0.31	4.92±0.22	4.27±0.30	37
flux growth	3.45±0.96	3.64±1.14	3.59±1.33	2.88±1.63	2.74±1.47	4.24±0.69	4.97±0.13	3.64±0.94	29
chemical bath deposition	3.39±0.80	3.66±0.95	3.29±1.11	2.70±1.65	2.41±1.47	4.02±0.70	4.80±0.37	3.47±0.85	28
template-directed	3.75±0.71	3.96±0.85	3.94±0.85	3.29±1.55	3.27±1.41	4.33±0.41	4.90±0.25	3.92±0.75	24
atomic layer deposition	4.10±0.33	4.25±0.77	4.27±0.57	4.17±0.83	4.17±0.41	4.60±0.36	4.88±0.22	4.36±0.34	24
combustion	4.46±0.41	4.40±0.51	4.75±0.39	4.42±0.97	4.77±0.39	4.71±0.39	4.94±0.17	4.63±0.28	24
float zone & Bridgman	3.73±0.78	3.70±1.32	3.52±1.41	3.95±1.25	2.91±1.41	4.32±0.66	4.91±0.29	3.87±0.84	22
microwave-assisted	4.05±0.28	4.25±0.53	4.32±0.44	4.10±0.45	3.80±0.62	4.58±0.24	4.72±0.38	4.26±0.29	20
lithographic patterning	3.87±0.64	4.17±0.49	4.03±0.67	4.10±0.57	3.67±0.79	4.43±0.53	4.93±0.18	4.18±0.45	15
coprecipitation	4.21±0.33	4.12±0.43	4.50±0.37	3.83±0.83	4.08±0.19	4.62±0.31	4.92±0.19	4.32±0.25	12
spark plasma sintering	4.00±0.00	4.05±0.27	4.23±0.26	4.32±0.34	4.14±0.23	4.45±0.15	4.95±0.15	4.32±0.12	11
mechanochemical	3.94±0.88	3.94±0.88	3.89±1.34	4.11±1.27	3.89±1.11	4.44±0.53	4.78±0.36	4.14±0.79	9
sonochemical	4.08±0.20	4.25±0.27	4.25±0.27	4.00±0.00	3.83±0.26	4.50±0.00	5.00±0.00	4.28±0.08	6
spray pyrolysis	4.33±0.41	4.58±0.38	4.50±0.55	4.67±0.41	4.25±0.76	4.75±0.27	5.00±0.00	4.58±0.33	6
electrospinning	4.00±0.00	4.00±0.00	4.38±0.25	4.00±0.00	4.12±0.25	4.38±0.25	5.00±0.00	4.28±0.13	4
ion implantation	3.83±0.29	4.00±0.00	4.50±0.50	3.83±1.61	3.67±0.58	4.50±0.50	5.00±0.00	4.20±0.53	3
liquid-phase epitaxy	4.00±nan	4.00±nan	4.00±nan	2.00±nan	4.00±nan	4.50±nan	5.00±nan	3.90±nan	1
incipient wetness impregnation	4.00±nan	4.00±nan	4.00±nan	4.00±nan	4.00±nan	4.50±nan	5.00±nan	4.20±nan	1
arc discharge	-	-	-	-	-	-	-	-	0

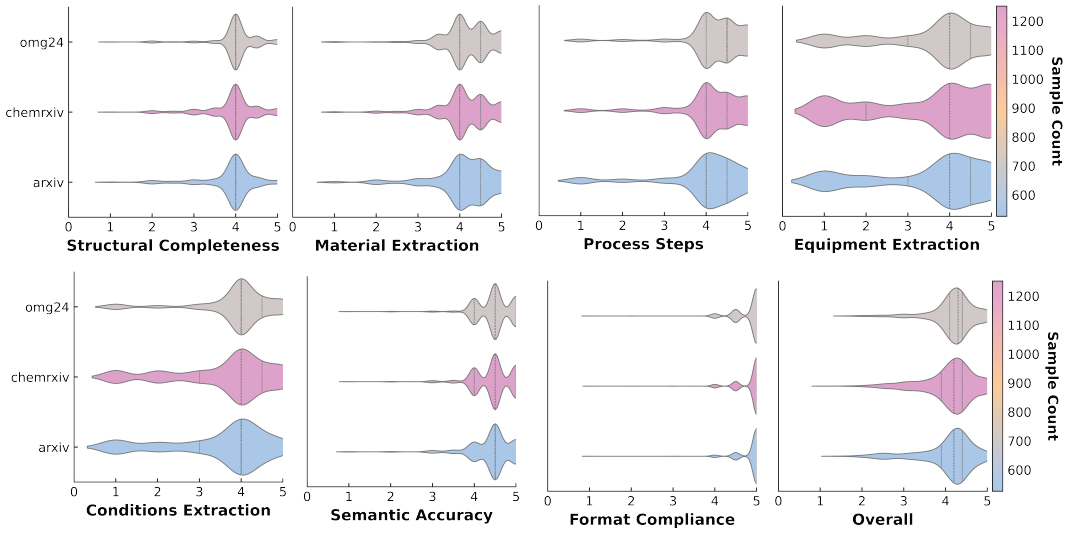

Figure 13: Distribution of LLM-judged overall extraction scores across different sources from `LeMat-Synth`. Each violin plot shows the probability density of the scores for a given synthesis type.

Table 6: Average LLM-judged extraction scores for the most frequent material types in the evaluated dataset subset ($N = 2483$ procedures). Scores are reported as mean $\pm$ standard deviation on a 1–5 scale. The Overall Score is the average of all seven evaluation criteria.

Material category	Structural completeness	Material completeness	Process steps	Equipment extraction	Condition extraction	Semantic accuracy	Format compliance	Overall score	Count
nanomaterials	4.01±0.47	4.14±0.57	4.21±0.68	3.65±1.24	3.76±0.97	4.48±0.41	4.85±0.29	4.16±0.51	476
framework & porous materials	3.95±0.57	4.15±0.67	4.12±0.84	3.45±1.47	3.63±1.19	4.50±0.43	4.88±0.30	4.09±0.61	385
functional materials & catalysts	3.93±0.61	4.14±0.63	4.12±0.76	3.32±1.51	3.52±1.21	4.44±0.45	4.88±0.26	4.05±0.61	351
ceramics & glasses	3.94±0.65	4.10±0.77	4.07±0.95	3.80±1.32	3.83±1.15	4.43±0.53	4.90±0.26	4.15±0.67	270
semiconductors & electronic	3.95±0.57	4.16±0.64	4.13±0.84	3.64±1.31	3.60±1.16	4.48±0.42	4.90±0.23	4.13±0.58	255
composites	4.06±0.35	4.23±0.41	4.27±0.54	3.79±0.97	3.90±0.68	4.51±0.34	4.86±0.26	4.23±0.35	154
other	3.75±0.99	4.20±0.69	3.88±1.26	3.26±1.61	3.59±1.36	4.33±0.87	4.71±0.76	3.96±0.89	152
polymers & soft matter	3.96±0.50	4.13±0.61	4.20±0.68	3.43±1.38	3.62±1.05	4.42±0.42	4.84±0.29	4.08±0.54	132
metals & alloys	3.99±0.45	4.11±0.75	4.23±0.66	3.87±1.21	3.78±1.01	4.48±0.49	4.89±0.31	4.19±0.51	92
two-dimensional materials	3.88±0.71	4.10±0.63	4.05±1.07	3.52±1.30	3.56±1.10	4.39±0.49	4.90±0.24	4.06±0.66	89
biomaterials & biological	3.77±0.60	4.01±0.62	4.02±0.69	3.48±1.59	3.49±1.25	4.40±0.40	4.85±0.30	4.00±0.60	66
hybrid & organic-inorganic	3.93±0.64	4.02±0.70	4.25±0.77	3.49±1.50	3.71±1.23	4.44±0.38	4.86±0.28	4.10±0.65	51
energy & sustainability	4.31±0.65	4.50±0.46	4.50±0.46	4.12±1.33	4.19±0.65	4.69±0.37	4.88±0.23	4.45±0.45	8
emerging & quantum materials	4.50±0.71	4.50±0.71	4.75±0.35	4.50±0.71	4.50±0.71	4.75±0.35	4.75±0.35	4.60±0.57	2
liquid materials	-	-	-	-	-	-	-	-	0

A.3 Figure extraction

Segmenting large figures into sub-plots. To extract individual subplots from figures in research papers, we employ the DINO model [29] with zero-shot image segmentation. The prompt 'a plot' is used to guide the model in localizing subplot regions, with both text and box confidence thresholds set to 0.3. After initial detection, a post-processing step refines the bounding boxes to ensure complete coverage of each subplot, including axis labels and tick marks. To distinguish multi-panel figures from single-plot figures, we retain only bounding boxes that cover less than 50% of the total figure area; larger boxes are assumed to correspond to entire figures and are excluded. Empirical results indicate that this approach reliably identifies subplots across a variety of figure types.

Classifying plots with quantitative data. To classify segmented subplots and full-figure plots, we employ a ResNet-152 model [68], pretrained on ImageNet and fine-tuned on the DocFig dataset [30]. The dataset is split into 19,000 samples for training and 13,000 samples for testing. The model is trained with default hyperparameters for 20 epochs using the Adam optimizer with a learning rate of 1e-3. Our classification task focuses exclusively on the plot types "line chart", "bar plot" and "scatter plot" which are relevant for downstream information extraction; qualitative figures are excluded from further processing. The fine-tuned model achieves an F1-score of 88.03% on the test set, indicating strong performance in accurately identifying quantitative plots for subsequent analysis.

Extracting data with a vision LLM. To convert these numerical figures into a structured and interpretable format for further use, we explore the capabilities of advanced vision-language models to extract data from line plots, focusing on 2D coordinate retrieval. Inspired by [33], where multimodal models were used to extract and regenerate plots, we use `Claude-Sonnet-4` (`claude-sonnet-4-20250514`) to extract 2D coordinates with their corresponding series names, as well as metadata fields like titles, axis labels, and units. The model is prompted to output a JSON object in a predefined schema, which is then parsed into a Pydantic object to ensure data consistency and structured integration into our data extraction pipeline.

A.3.1 Figure Extraction Evaluation

Manual annotations. For each series, the extracted coordinates are matched to the closest ground truth points using nearest-neighbor matching. This matching is performed in a normalized coordinate space, where both x and y axes are scaled to their respective ranges to ensure that errors are comparable across axes. The normalization scale is computed from the minimum and maximum values of the ground truth coordinates for each axis. We manually annotate 15 line charts from selected papers in catalysis [69, 70, 71, 72, 73]. For expanding the pipeline in the future, we plan to annotate larger samples from a more diverse array of plot types, e.g. scatter, bar and box plots.

The evaluation is based on two error metrics:

- **Root Mean Square Error (RMSE):** which penalizes larger errors more heavily due to its quadratic nature.
- **Mean Absolute Error (MAE):** which treats all deviations linearly, providing a robust average error.

To compute the error metrics for a single series, we define the extracted points as:

$$\mathcal{P} = \{(x_i, y_i) \,|\, i \in \{1, \ldots, N\}\} \tag{1}$$

and the ground truth points as:

$$\mathcal{G} = \left\{(x_j^*, y_j^*) \,\middle|\, j \in \{1, \ldots, M\}\right\} \tag{2}$$

Compute the normalization scales for each axis as:

$$S_x = \max_j x_j^* - \min_j x_j^*, \quad S_y = \max_j y_j^* - \min_j y_j^* \tag{3}$$

For each extracted point (x_i, y_i), we find the nearest ground truth point by computing the normalized Euclidean distance:

$$d_i = \min_j \sqrt{\left(\frac{x_i - x_j^*}{S_x}\right)^2 + \left(\frac{y_i - y_j^*}{S_y}\right)^2} \tag{4}$$

The RMSE is then defined as:

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} d_i^2} \tag{5}$$

and the MAE as:

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N} d_i \tag{6}$$

A.4 Prompts

This section shows the system prompts employed and the full configurations used (incl. signatures and LLM configurations) to extract the data presented in this work.

Filtering papers

Prompt

```
Analyze the following text and answer the questions in JSON format:
{chunk}
Questions:
1. Does it contain a material synthesis recipe?
    (Answer with true or false)
2. If yes, what is the material name?
    (Answer with the material name or "N/A" if no recipe)
3. If yes, which category of materials does it belong to?
    (Answer with the specific material type or "N/A" if no recipe)
List of material categories:
Metals, Ceramics, Semiconductors, Superconductors, Composites,
Biomaterials, Nanomaterials, Polymers, Magnetic, Textiles, Chemicals, Other
Format your response as a JSON object with the following structure:
{{
"contains_recipe": true/false,
"material_name": "material name or N/A",
"material_category": "material category or N/A"
}}
```

Material extraction

Prompt

```
You are a helpful assistant that extracts ONLY the final synthesized materials
↪  from scientific papers.

Your task is to identify ONLY the materials that are the final products of
↪  synthesis procedures described in the paper.

IMPORTANT GUIDELINES:
- ONLY include materials that are the final synthesized products
- DO NOT include starting materials, precursors, supports, gases, solvents, or
↪  other chemicals used in synthesis
- DO NOT include materials that are just mentioned or characterized but not
↪  synthesized
- Focus on the main target materials that are actually synthesized

EXAMPLES OF WHAT TO INCLUDE:
- "Ni/Al2O3" (if Ni is deposited on Al2O3)
- "Ir/SiO2" (if Ir is supported on SiO2)
- "LiFePO4 nanoparticles" (if LiFePO4 is synthesized)
- "Co-doped LiFePO4" (if this specific material is synthesized)

EXAMPLES OF WHAT TO EXCLUDE:
- "Ni", "Ir", "Ru" (if these are just precursors)
- "H-ZSM-5", "Al2O3", "SiO2" (if these are just supports)
- "Ammonia", "Argon", "Hydrogen" (gases)
- "Deionized water" (solvents)
- "Ammonium hydroxide" (reagents)

Return a simple comma-separated list of ONLY the final synthesized materials.

If no materials are synthesized in the paper, return "No materials
↪  synthesized".

Keep the output simple and clean - just the final synthesized material names
↪  separated by commas.
```

Configuration (YAML)

```yaml
architecture:
  _target_: llm_synthesis.transformers.material_extraction.dspy_extraction.Dsp
  ↪  yTextExtractor
  signature:
    _target_: llm_synthesis.transformers.material_extraction.dspy_extraction.m
    ↪  ake_dspy_text_extractor_signature
    signature_name: "TextToMaterials"
    instructions: "Extract ONLY the final synthesized materials from the
    ↪  publication text."
    input_description: "The publication text to extract the final synthesized
    ↪  materials from."
    output_name: "materials"
    output_description: "The final synthesized materials as a comma-separated
    ↪  list."
  lm:
    _target_: llm_synthesis.utils.dspy_utils.get_llm_from_name
    llm_name: "gemini-2.0-flash"
    model_kwargs:
      temperature: 0.0
    system_prompt:
```

```
    _target_: llm_synthesis.utils.read_prompt_str_from_txt
    prompt_path: "examples/system_prompts/material_extraction/default.txt"
```

Synthesis extraction

Prompt

```
You are a helpful assistant that extracts the structured synthesis for a specific material from the paper text.

Focus ONLY on the synthesis procedure for the specified material. Search through the entire paper text to find the synthesis procedure that describes how this specific material is made.

IMPORTANT: You must output ONLY a valid JSON object with a "structured_synthesis" field. Do not include any reasoning, explanations, or markdown formatting.

If you cannot find a synthesis procedure for the specified material, return a minimal structure with the material name and an empty synthesis.

The JSON output must follow this exact structure:
{
  "structured_synthesis": {
    "target_compound": "string (required) - should match the specified material name",
    "target_compound_type": "string (required) - choose from: 'metals & alloys', 'ceramics & glasses', 'polymers & soft matter', 'composites', 'semiconductors & electronic', 'nanomaterials', 'two-dimensional materials', 'framework & porous materials', 'biomaterials & biological', 'liquid materials', 'hybrid & organic-inorganic', 'functional materials', 'energy & sustainability', 'smart & responsive materials', 'emerging & quantum materials', 'other'",
    "synthesis_method": "string (required) - choose from: 'PVD', 'CVD', 'arc discharge', 'ball milling', 'spray pyrolysis', 'electrospinning', 'sol-gel', 'hydrothermal', 'solvothermal', 'precipitation', 'coprecipitation', 'combustion', 'microwave-assisted', 'sonochemical', 'template-directed', 'solid-state', 'flux growth', 'float zone & Bridgman', 'arc melting & induction melting', 'spark plasma sintering', 'electrochemical deposition', 'chemical bath deposition', 'liquid-phase epitaxy', 'self-assembly', 'atomic layer deposition', 'molecular beam epitaxy', 'pulsed laser deposition', 'ion implantation', 'lithographic patterning', 'wet impregnation', 'incipient wetness impregnation', 'mechanical mixing', 'other'
    "starting_materials": [{"name": "string", "amount": "number or null", "unit": "string or null", "purity": "string or null", "vendor": "string or null"}],
    "steps": [{"step_number": "integer", "action": "string", "description": "string or null", "materials": [{"name": "string", "amount": "number or null", "unit": "string or null", "purity": "string or null", "vendor": "string or null"}], "equipment": [{"name": "string", "instrument_vendor": "string or null", "settings": "string or null"}], "conditions": {"temperature": "number or null", "temp_unit": "string or null", "duration": "number or null", "time_unit": "string or null", "pressure": "number or null", "pressure_unit": "string or null", "atmosphere": "string or null", "stirring": "boolean or null", "stirring_speed": "number or null", "ph": "number or null"}}],
    "equipment": [{"name": "string", "instrument_vendor": "string or null", "settings": "string or null"}],
    "notes": "string or null"
```

```
  }
}

Do not include any text before or after the JSON object. Output only the JSON.
```

Configuration (YAML)

```yaml
architecture:
  _target_: llm_synthesis.transformers.synthesis_extraction.dspy_synthesis_ext
  ↪  raction.DspySynthesisExtractor
  signature:
    _target_: llm_synthesis.transformers.synthesis_extraction.dspy_synthesis_e
    ↪  xtraction.make_dspy_synthesis_extractor_signature
    signature_name: "SynthesisSignature"
    instructions: "Extract the structured synthesis for a specific material
    ↪  from the paper text."
    paper_text_description: "The complete paper text to search for the
    ↪  material's synthesis procedure."
    material_name_description: "The name of the specific material to extract
    ↪  synthesis for."
    output_name: "structured_synthesis"
    output_description: "The extracted structured synthesis for the specific
    ↪  material."
  lm:
    _target_: llm_synthesis.utils.dspy_utils.get_llm_from_name
    llm_name: "gemini-2.0-flash"
    model_kwargs:
      temperature: 0.0
      max_tokens: 8000
      max_retries: 3
    system_prompt:
      _target_: llm_synthesis.utils.read_prompt_str_from_txt
      prompt_path: "examples/system_prompts/synthesis_extraction/default.txt"
```

Figure extraction

For figure extraction, we do not provide a separate DSPy configuration. Unlike material and synthesis extraction (which are wrapped with DSPy signatures and explicit input/output schemas), the figure extraction pipeline directly leverages the system prompt together with a Claude API client. In this setup, the model is invoked with the raw prompt and image data, and the parsing into structured objects (`ExtractedLinePlotData`) happens entirely within the custom transformer implementation. Because no DSPy signature or schema mediation is involved, there is no corresponding YAML configuration block to display. Instead, the logic is captured in the prompt (shown below) and the Python implementation excerpted below.

Prompt

```
LINE_CHART_PROMPT = """
You will be provided with a line chart. The chart may not be chunked very well,
so you may need to read only the plot in the center of the image.
In the chart, there will be several lines representing different data series.

1. Identify the different lines by their colors and labels.
2. For each line, extract the coordinates of the points that make up the line.
   Do not include any points that are not part of the line.
3. If the chart has metadata such as a title, x-axis label, y-axis labels,
   or units, extract that information as well.
   Keep the scientific terms in Markdown format.
4. Output the data in the specified format:

Name_of_Line_1: [[x1, y1], [x2, y2], ...]
```

```
title:
x_axis_label:
x_axis_unit:
y_left_axis_label:
y_left_axis_unit:

Do not output any other text, just the data in the format above.
"""
```

Implementation excerpt (Python)

```python
class ClaudeLinePlotDataExtractor(LinePlotDataExtractorInterface):
    def __init__(self, model_name: str,
                 prompt: str = resources.LINE_CHART_PROMPT,
                 max_tokens: int = 1024,
                 temperature: float = 0.0):
        super().__init__()
        self.claude_client = ClaudeAPIClient(model_name)
        self.prompt = prompt
        self.max_tokens = max_tokens
        self.temperature = temperature

    def forward(self, input: FigureInfoWithPaper) -> ExtractedLinePlotData:
        figure_base64 = input.base64_data
        self.claude_client.reset_cost()
        claude_response_obj = self.claude_client.vision_model_api_call(
            figure_base64=figure_base64,
            prompt=self.prompt,
            max_tokens=self.max_tokens,
            temperature=self.temperature,
        )
        return self._parse_into_pydantic(claude_response_obj)

    def _parse_into_pydantic(self, response: str) -> ExtractedLinePlotData:
        """Parse text into Pydantic object with regex pattern matching"""
        ...
```

Synthesis evaluation

In this case, the evaluation logic is fully captured within the DSPy configuration itself, so we do not provide a standalone prompt block. Both the task instructions and the system prompt are directly embedded inside the configuration file rather than stored separately. The complete configuration is shown below:

Configuration (YAML)

```yaml
architecture:
  _target_: llm_synthesis.metrics.judge.general_synthesis_judge.DspyGeneralSyn⌋
  ↪  thesisJudge
  signature:
    _target_: llm_synthesis.metrics.judge.general_synthesis_judge.make_general⌋
    ↪  _synthesis_judge_signature
    signature_name: "GeneralSynthesisJudgeSignature"
    instructions: >
      You are an expert materials scientist and data extraction specialist with
      ↪  extensive experience in:
        - Synthesis procedure analysis and documentation
        - Structured data extraction from scientific literature
        - Materials science ontology design and terminology standardization
        - Quality assessment of automated scientific information extraction
        ↪  systems

      Evaluate how well the GeneralSynthesisOntology extraction captures
      ↪  synthesis information from
      the provided source text.
```

```
    IMPORTANT: Do NOT penalize the extraction system for failing to include
    ↪  information that is
    not present in the original paper. Missing elements should only be
    ↪  considered errors if they
    were clearly stated in the source but were not extracted. If an element
    ↪  is absent in both the
    source and the extraction, and is correctly left blank or omitted, this
    ↪  should be considered
    correct and scored highly.

    ASSESSMENT FOCUS:
      - Completeness: All synthesis components present in the source are
      ↪  captured
      - Accuracy: Correct values, units, and classifications based on the
      ↪  text
      - Structure: Proper organization and logical sequencing of elements
      - Semantic Preservation: Scientific meaning and intent faithfully
      ↪  maintained
      - Schema Compliance: Conforms to the expected ontology format and data
      ↪  types

    EVALUATION CRITERIA (Score 1-5 for each):
      1. Structural Completeness - Extraction of all relevant synthesis
      ↪  components from the source (materials, steps, equipment,
      ↪  conditions)
      2. Material Extraction - Correct names, quantities, units, purities as
      ↪  specified in the paper
      3. Process Steps - Accurate step order and correct action
      ↪  classification
      4. Equipment Extraction - Proper identification of all equipment
      ↪  explicitly mentioned
      5. Conditions Extraction - Accurate recording of parameters such as
      ↪  temperature, time, atmosphere, pressure, etc.
      6. Semantic Accuracy - Faithful preservation of scientific meaning
      ↪  without misinterpretation
      7. Format Compliance - Adherence to ontology schema, data types, and
      ↪  field structure

    For each criterion:
      - Assign a score between 1 and 5
      - Provide detailed technical reasoning for the assigned score
      - Offer specific, constructive recommendations for improvement, if
      ↪  applicable

lm:
  _target_: llm_synthesis.utils.dspy_utils.get_llm_from_name
  llm_name: "gemini-2.0-flash"
  model_kwargs:
    temperature: 0.1
    max_tokens: 4096
  system_prompt: >
    You are a senior materials scientist and data extraction expert with deep
    ↪  expertise in:
      - Inorganic and organic synthesis methodologies
      - Laboratory instrumentation and experimental workflows
      - Chemical nomenclature, stoichiometry, and unit conventions
      - Optimization of synthesis conditions and reaction parameters
      - Structured data modeling and materials science ontology design
      - Evaluation methodologies for automated information extraction systems

    Your assessments should reflect best practices in synthesis reporting and
    ↪  uphold the highest
```

```
    standards of scientific accuracy, reproducibility, and structured data
    ↪  quality.

    When evaluating extracted synthesis data:
      - Rely on your domain expertise to assess technical correctness,
      ↪  semantic fidelity, and structural organization
      - Emphasize clarity, precision, and alignment with real-world
      ↪  experimental protocols
      - Consider the intended schema and use context to assess compliance and
      ↪  completeness
      - Do not penalize the extraction system for omitting elements that were
      ↪  not explicitly present in the source text

    Your evaluation should be technically rigorous, yet fair, grounded in
    ↪  both materials science principles and data extraction best practices.

enable_reasoning_traces: true
confidence_threshold: 0.7
```