Supplementary Information for Expert-Grounded Automatic Prompt Engineering for Extracting Lattice Constants of High-Entropy Alloys from Scientific Publications using Large Language Models

Shunshun Liu,[1] Talon R. Booth,[1] Yangfeng Ji,[2]
Wesley Reinhart,[3,4] and Prasanna V. Balachandran[1,*]

[1]*Department of Materials Science and Engineering, University of Virginia, Charlottesville, VA, USA*

[2]*Department of Computer Science, University of Virginia, Charlottesville, VA, USA*

[3]*Department of Materials Science and Engineering, Pennsylvania State University, State College, PA, USA*

[4]*Institute for Computational and Data Sciences, Pennsylvania State University, State College, PA, USA*

(Dated: November 20, 2025)

* pvb5e@virginia.edu

S1. PROMPTS USED FOR CONSTRUCTING THE LLM APPLICATION SURVEY FIGURE

The Figure S1 was created using data from the Ai2 Asta tool [1]. The following search terms were used:

- Find papers that use LLM in materials science
- Work harder
- Use the following categories: Knowledge Extraction & Curation, Property Prediction & Modeling, Materials & Molecular Design (Generative), Scientific Reasoning & Hypothesis Generation, Workflow Automation & Self-Driving Laboratories, Data Management & Knowledge Infrastructure, Education & Scientific Communication, can you count how many papers in each class?

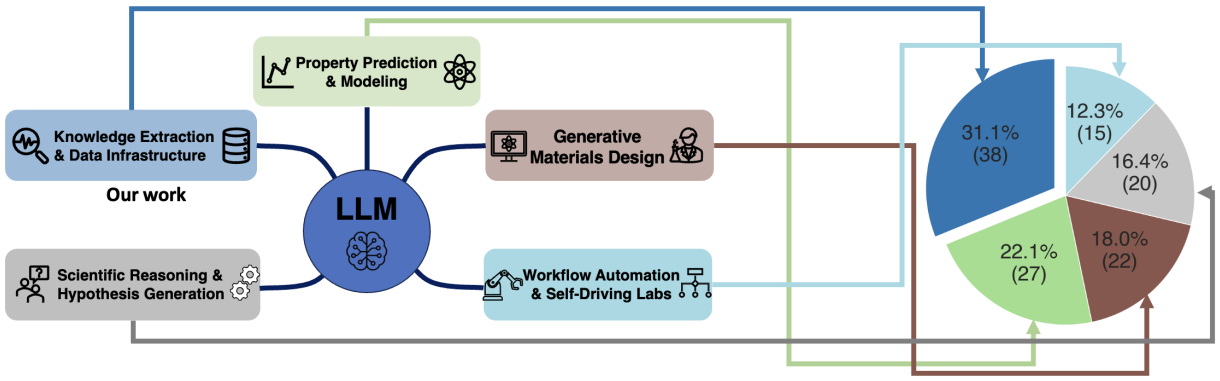

Figure S1: Applications of LLMs in materials science. Survey of 122 publications across five major application categories produced the following results. Knowledge Extraction and Data Infrastructure (31.1%) represents the most common use, involving automated knowledge and data extraction from scientific literature. Property Prediction and Modeling (22.1%) includes approaches for structure-property relationships. Generative Materials Design (18.0%) focuses on inverse design and computational generation of novel materials. Workflow Automation and Self-Driving Labs (16.4%) covers autonomous computational and experimental materials science. Scientific Reasoning and Hypothesis Generation (12.3%) addresses mechanistic understanding and identification of promising research directions. Source: Ai2-Asta [1].

S2. PROMPTS USED IN THE PAPER FOR DATA EXTRACTION

```
Read the PDF file and extract required parameters.
You will ONLY report parameters of single phase high entropy alloy.
Lattice constant should be in the unit of angstrom, if lattice constant does not present, only report *None*.
The output MUST be in json format, follow the schema example:
{
  "name": "Al0.25NbTiMoV"
  "nominal_composition": "Al0.25Nb1.0Ti1.0Mo1.0V1.0",
  "measured_composition": "Al5.4Ti22.0V22.1Nb25.5Mo25.0",
  "lattice_constant": 3.206,
  "phases": "BCC",
  "alloy_condition": "As-cast",
  "doi": "10.3390/e16020870"
}
```

Figure S2: Initial prompt for data extraction designed by the authors without any consultation with the LLM.

```
Read the PDF file and extract the following parameters for all high-entropy alloys discussed in the results section:

1. name (string, e.g., "AlCoCrFeNi")
2. nominal_composition (string, representing the stoichiometric ratio of each element, e.g., "Al1.0Co1.0Cr1.0Fe1.0Ni1.0". If an element's ratio is not explicitly stated, assume 1.0)
3. measured_composition (string, exactly as written in the paper)
4. lattice_constant (float, in angstroms, rounded to 3 decimal places)
5. phases (string, e.g., "BCC")
6. alloy_condition (string, e.g., "As-Cast")
7. doi (string)

Extract parameters primarily from the text. Use data from tables only if the text data is incomplete. Use figures as a last resort. If a parameter is truly missing from the PDF for a given alloy, explicitly report it as "Not found" rather than omitting it.

For each parameter, include a confidence score (0-100) indicating your certainty in the extracted information. Consider a score of 90 or above as high confidence.

The output should be a list of JSON objects, one for each alloy discussed in the paper, in the following format:

[
    {
        "name": "AlloyName",
```

```
        "nominal_composition": "Element11.0Element21.0...",
        "measured_composition": "Composition as written",
        "lattice_constant": X.XXX,
        "phases": "Phase1,Phase2,...",
        "alloy_condition": "condition",
        "doi": "DOI",
        "confidence_scores": {
            "name": XX,
            "nominal_composition": XX,
            "measured_composition": XX,
            "lattice_constant": XX,
            "phases": XX,
            "alloy_condition": XX,
            "doi": XX
        }
    },
    ...
]

Include an alloy in the output only if it is explicitly discussed in the results section. Ensure that the output format and data closely match the provided schema. If information for a specific parameter is not available, use "Not found" and assign a low confidence score.

Example of correct output:
[
    {
        "name": "HfNbTaTiZr",
        "nominal_composition": "Hf1.0Nb1.0Ta1.0Ti1.0Zr1.0",
        "measured_composition": "Hf20.8Nb18.9Ta20.2Ti20.2Zr19.9",
```

```
        "lattice_constant": 3.414,
        "phases": "BCC",
        "alloy_condition": "As-Cast",
        "doi": "10.1016/j.jallcom.2014.11.064",
        "confidence_scores": {
            "name": 100,
            "nominal_composition": 90,
            "measured_composition": 95,
            "lattice_constant": 100,
            "phases": 100,
            "alloy_condition": 95,
            "doi": 100
        }
    }
]

Example of output with missing data:
[
    {
        "name": "AlCoCrFeNi",
        "nominal_composition": "Al1.0Co1.0Cr1.0Fe1.0Ni1.0",
        "measured_composition": "Not found",
        "lattice_constant": 3.567,
        "phases": "FCC",
        "alloy_condition": "Not found",
        "doi": "10.1016/j.example.2023.01.001",
        "confidence_scores": {
            "name": 100,
            "nominal_composition": 90,
            "measured_composition": 0,
```

```
            "lattice_constant": 95,
            "phases": 100,
            "alloy_condition": 0,
            "doi": 100
        }
    }
]
```

Figure S3: Optimized prompt from the TextGrad output.

S3. TEXT LOSS RETURNED FOR INDIVIDUAL PAPERS

```
Based on the evaluation criteria, here is my assessment of the output:
1. Does the output match the format provided in the question?
Yes, the output is a list of JSON objects matching the specified format.
2. Are the extracted materials high-entropy alloys?
No, these are not high-entropy alloys. They are ferritic alloys with G-phase precipitates.
3. Are the alloy compositions in the output correct and reflect the material?
Yes, the nominal compositions match those given in Table 1 of the paper.
4. Are the lattice constants in the output formatted in angstrom?
Yes, the lattice constants are given in angstroms (11.22 Å).
5. Are the lattice constants in the output truly the lattice constant of the material not others?
Yes, 11.22 Å matches the lattice parameter given for the Ni16Ti6Si7-G phase in Table 2.
6. Are the phase information in the output correct and reflect the material?
Yes, the phases listed (BCC, G-phase, L21) match those discussed in the paper.
7. Are the alloy conditions in the output correct and reflect the material?
Yes, the conditions listed (As-Cast, Solution-Treated, Cold-Rolled) are mentioned in the paper.
8. Is the DOI in the output matches with the original paper?
Yes, the DOI matches the one given at the top of the paper.
9. Is there any HEA high-entropy alloys missed in the provided output?
No, there are no high-entropy alloys discussed in this paper, so none are missed.
```

Figure S4: Text loss returned for Yang et al. [2] using the `EvaluatePdfOutputLoss` function.

```
I will evaluate the output according to the provided criteria:
1. Does the output match the format provided in the question?
Yes
2. Are the extracted materials high-entropy alloys?
No, the extracted alloy is described as a medium entropy alloy (MEA).
3. Are the alloy compositions in the output correct and reflect the material?
Yes
4. Are the lattice constants in the output formatted in angstrom?
Yes
5. Are the lattice constants in the output truly the lattice constant of the material not others?
No, the reported 10.9 Å is the lattice parameter of the Cr23C6 carbide, not the MEA matrix.
6. Are the phase information in the output correct and reflect the material?
Yes
7. Are the alloy conditions in the output correct and reflect the material?
Yes
8. Is the DOI in the output matches with the original paper?
Yes
9. Is there any HEA high-entropy alloys missed in the provided output?
No, the paper focuses on a medium entropy alloy, not a high entropy alloy.
```

Figure S5: Text loss returned for Pan et al. [3] using the `EvaluatePdfOutputLoss` function.

```
I will evaluate the output according to the provided criteria:
1. Does the output match the format provided in the question?
Yes, the output matches the JSON format specified in the question.
2. Are the extracted materials high-entropy alloys?
Yes, all extracted materials are high-entropy alloys from the FeCoCrAlNiTix system.
3. Are the alloy compositions in the output correct and reflect the material?
Yes, the nominal and measured compositions appear to correctly reflect the materials described in the paper.
4. Are the lattice constants in the output formatted in angstrom?
The lattice constant is only provided for FeCoCrAlNiTi0.5 and is in angstroms. For the others it's listed as "Not found".
5. Are the lattice constants in the output truly the lattice constant of the material not others?
The one lattice constant provided (11.303 Å for FeCoCrAlNiTi0.5) appears to be for the Ti2Ni phase rather than the overall alloy lattice constant.
6. Are the phase information in the output correct and reflect the material?
Yes, the phase information appears to be correct based on the description in the paper.
7. Are the alloy conditions in the output correct and reflect the material?
Yes, "As-Cast" is correct for these laser surface alloyed coatings.
8. Is the DOI in the output matches with the original paper?
Yes, the DOI matches the one given in the paper.
9. Is there any HEA high-entropy alloys missed in the provided output?
No, all FeCoCrAlNiTix alloys discussed in the results section appear to be included.
```

Figure S6: Text loss returned for Wu et al. [4] using the `EvaluatePdfOutputLoss` function.

```
Here is my evaluation of the output:
1. Does the output match the format provided in the question? Yes
2. Are the extracted materials high-entropy alloys? No, this is described as a medium-entropy alloy.
3. Are the alloy compositions in the output correct and reflect the material? Yes, the nominal composition matches what is stated in the paper.
4. Are the lattice constants in the output formatted in angstrom? Yes
5. Are the lattice constants in the output truly the lattice constant of the material not others? No, the value given (0.255) appears to be the Burgers vector magnitude, not the lattice constant.
6. Are the phase information in the output correct and reflect the material? Yes, the phases listed match those described in the paper.
7. Are the alloy conditions in the output correct and reflect the material? Yes, the annealed condition is mentioned in the paper.
8. Is the DOI in the output matches with the original paper? Yes
9. Is there any HEA high-entropy alloys missed in the provided output? No, the paper focuses on this single medium-entropy alloy.
```

Figure S7: Text loss returned for Moon et al. [5] using the `EvaluatePdfOutputLoss` function.

```
I'll evaluate the output according to the provided criteria:
1. Does the output match the format provided in the question?
Yes, the output is a list of JSON objects with the specified fields.
2. Are the extracted materials high-entropy alloys?
Yes, both NiCuPdTiZrHf and NiCuTiZrHf are high-entropy alloys.
3. Are the alloy compositions in the output correct and reflect the material?
Yes, the nominal compositions match the information given in the paper.
4. Are the lattice constants in the output formatted in angstrom?
Yes, the lattice constant for NiCuPdTiZrHf is given in angstroms (0.319 Å).
5. Are the lattice constants in the output truly the lattice constant of the
↪ material not others?
Yes, the 0.319 Å value matches the B2 lattice constant reported in the paper for
↪ NiCuPdTiZrHf.
6. Are the phase information in the output correct and reflect the material?
Yes, the phases listed (B2, B19') match those discussed in the paper for these
↪ alloys.
7. Are the alloy conditions in the output correct and reflect the material?
Yes, the as-cast and homogenized conditions are correctly noted.
8. Is the DOI in the output matches with the original paper?
Yes, the DOI matches the one given in the paper.
9. Is there any HEA high-entropy alloys missed in the provided output?
No, the output covers the two main HEAs discussed in detail in the results
↪ section.
```

Figure S8: Text loss returned for Piorunek et al. [6] using the `EvaluatePdfOutputLoss` function.

```
I will evaluate the output according to the provided criteria:
1. Does the output match the format provided in the question?
Yes, the output matches the requested JSON format.
2. Are the extracted materials high-entropy alloys?
No, these are not high-entropy alloys. They are Mg-based alloys with Gd, Y, and Zr additions.
3. Are the alloy compositions in the output correct and reflect the material?
Yes, the measured compositions match the data provided in Table 1 of the paper.
4. Are the lattice constants in the output formatted in angstrom?
Yes, the lattice constants are given in angstroms (e.g. 0.323 Å).
5. Are the lattice constants in the output truly the lattice constant of the material not others?
Yes, the lattice constants match the 'a' parameter values given in Table 3 of the paper.
6. Are the phase information in the output correct and reflect the material?
Yes, the phase information matches what is described in the paper for each composition.
7. Are the alloy conditions in the output correct and reflect the material?
Yes, the alloy condition "As-deposited LDED" correctly reflects the laser directed energy deposition process used.
8. Is the DOI in the output matches with the original paper?
Yes, the DOI matches the one given in the paper.
9. Is there any HEA high-entropy alloys missed in the provided output?
No, there are no high-entropy alloys in this paper, so none are missed.
```

Figure S9: Text loss returned for Jiang et al. [7] using the `EvaluatePdfOutputLoss` function.

```
I'll evaluate the output for each extracted material according to the provided criteria:
1. Does the output match the format provided in the question?
Yes. The output is a list of JSON objects with the required fields.
2. Are the extracted materials high-entropy alloys?
Yes. All extracted alloys are high-entropy alloys from the CoFeNiMnTixAl1-x system.
3. Are the alloy compositions in the output correct and reflect the material?
Yes. The compositions match those given in the paper for each alloy.
4. Are the lattice constants value in the output formatted in angstrom?
Yes. The lattice constants are given in angstroms.
5. Are the lattice constants in the output truly the lattice constant of the material not others?
Yes. The lattice constants match those reported in the paper for the BCC phase.
6. Are the phase information in the output correct and reflect the material?
Yes. The phase information matches what is reported in the paper for each composition.
7. Are the alloy conditions in the output correct and reflect the material?
Yes. The alloys are reported as "As-Cast" which matches the paper.
8. Is the DOI in the output matches with the original paper?
Yes. The DOI matches the one given in the paper.
9. Is there any HEA high-entropy alloys missed in the provided output?
Yes. The output is missing some of the intermediate compositions discussed in the paper, such as Ti0.1Al0.9, Ti0.3Al0.7, and Ti0.8Al0.2.
```

Figure S10: Text loss returned for Cui et al. [8] using the `EvaluatePdfOutputLoss` function.

S4. PROMPT TRANSFERABILITY

TABLE S1: Comparison of entity-level extraction performance across Claude 4.5 Sonnet, Gemini 2.5 Flash, and GPT-5 LLM models using the *initial prompt* with different thinking or reasoning configurations. Claude 4.5 Sonnet was tested with and without extended reasoning. When the extended reasoning capability was enabled, the thinking budget was set at 1024 tokens. Extended reasoning is disabled by default for Claude 4.5 Sonnet. Gemini 2.5 Flash was evaluated with and without dynamic reasoning (dynamic reasoning is the default configuration). GPT-5 was tested with minimal and medium reasoning budgets (medium reasoning is the default configuration). All models were evaluated using their default temperature settings.

(a) Claude 4.5 Sonnet variants

Named Entity	**Claude 4.5 (extended reasoning)**			**Claude 4.5 (no reasoning)**		
	Precision	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
Nominal composition	0.909	1.000	0.952	0.667	1.000	0.800
Lattice constant	1.000	1.000	1.000	0.733	1.000	0.846
Phase	1.000	1.000	1.000	0.733	1.000	0.846
Alloy processing condition	1.000	1.000	1.000	0.733	1.000	0.846

(b) Gemini 2.5 Flash variants

Named Entity	**Gemini 2.5 Flash (dynamic reasoning)**			**Gemini 2.5 Flash (no reasoning)**		
	Precision	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
Nominal composition	1.000	0.818	0.900	0.654	0.773	0.708
Lattice constant	1.000	0.818	0.900	0.654	0.773	0.708
Phase	1.000	0.818	0.900	0.654	0.773	0.708
Alloy processing condition	1.000	0.818	0.900	0.654	0.773	0.708

(c) GPT-5 variants

Named Entity	**GPT-5 (medium reasoning)**			**GPT-5 (minimal reasoning)**		
	Precision	**Recall**	**F1 Score**	**Precision**	**Recall**	**F1 Score**
Nominal composition	1.000	0.864	0.927	0.850	0.895	0.872
Lattice constant	1.000	0.864	0.927	1.000	0.909	0.952
Phase	1.000	0.864	0.927	1.000	0.909	0.952
Alloy processing condition	1.000	0.864	0.927	1.000	0.909	0.952

TABLE S2: Comparison of data extraction performance across Claude 4.5 Sonnet, Gemini 2.5 Flash, and GPT-5 models using the *optimized prompt*. Precision and recall metrics were calculated with the same procedure as defined in Table 3. The best precision, recall, and F1 score is highlighted in bold font. The default configuration was used for all three LLM models: no reasoning for Claude 4.5 Sonnet, dynamic reasoning for Gemini 2.5 Flash and medium reasoning for GPT-5. The table also includes the cost (in USD) for processing the 46-publication set with each LLM, along with the cost-effectiveness ratio (average F1 score per dollar). Higher ratios indicate better extraction performance per unit cost.

(a) Claude 4.5 Sonnet no reasoning vs. Gemini 2.5 Flash dynamic reasoning vs. GPT-5 medium reasoning

Named Entity	**Claude 4.5 (no reasoning)**			**Gemini 2.5 Flash (dynamic reasoning)**			**GPT-5 (medium reasoning)**		
	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score
As-cast single-phase BCC	**1.000**	**0.977**	**0.988**	0.951	0.886	0.917	0.929	0.886	0.907
Nominal composition	**0.977**	**0.955**	**0.966**	0.951	0.886	0.917	0.905	0.864	0.884
Lattice constant	**0.953**	**0.932**	**0.942**	0.951	0.886	0.917	0.929	0.886	0.907
Approx. total cost (USD)		4.63			0.72			2.62	
Average F1 Score / USD		0.208			1.274			0.343	

(b) Claude 4.5 no reasoning vs. reasoning

Named Entity	**Claude 4.5 (no reasoning)**			**Claude 4.5 (reasoning)**		
	Precision	Recall	F1 Score	Precision	Recall	F1 Score
As-cast single-phase BCC	**1.000**	**0.977**	**0.988**	0.955	0.955	0.955
Nominal composition	**0.977**	**0.955**	**0.966**	0.955	0.955	0.955
Lattice constant	**0.953**	**0.932**	**0.942**	0.932	0.932	0.932
Approx. total cost (USD)		4.63			4.78	
Average F1 Score / USD		0.208			0.204	

S5. EXTENDED REASONING USING GEMINI 2.5 FLASH MODEL

```
**Initial Data Extraction and Analysis**
Alright, let's get down to business. My task is to meticulously extract specific information about high-entropy alloys (HEAs) from this document. I need to focus on the "Results and discussion" section, identifying each HEA and pulling out its name, nominal composition, measured composition, lattice constant, phases present, alloy condition, and the paper's DOI. I'll need to assign confidence scores to each piece of data, and if something's missing, I'll flag it as "Not found."
I'll start by systematically scanning the text. First, I'll hunt for HEA names. Then, for each HEA, I'll dig for the required parameters. The name is usually right there in the text. Nominal composition will be assumed equiatomic unless specified otherwise. Measured composition will be "Not found" unless precise percentages or ratios are given. Lattice constants will require finding Å values, ideally linked to diffraction data. Phases will be identified as "bcc," "fcc," etc. The alloy condition will be determined by looking for terms like "as-cast" or "annealed." Finally, the DOI is straightforward; it's the paper's unique identifier.
**Deep Dive into the Text**
Let's begin the search. I spot "TiVNbTa" and "TiVNbMoTa" right away in the first page's heading. That's a good start. I also see mentions of "bcc lattice" which tells me the phase of these alloys. Then, I found MoNbTaW and NiCoFeCrMn.  "arc melted" and "powder samples were ground from the as-cast ingots".
In the experimental details, it says the alloys were "arc melted" and that they confirmed a single "bcc structure".  And there's a specific lattice constant for TiVNbTa: 3.2206 (9) Å.
```

```
Now, into the "Results and discussion." TiVNbTa is being discussed a lot, and the lattice constant is repeated, so that is another confirmation of its bcc structure. TiVNbMoTa is also mentioned. I see comparisons between the two, which provides a lattice constant for TiVNbMoTa. Other alloys identified are "TiVNbZr", "TiVNbMo" (this needs careful attention, as it may or may not be the same as TiVNbMoTa. For now, I'm assuming it's a separate 4-element alloy), and more mentions of NiCoFeCr and NiCoFeCrPd (as comparisons). The general condition given is the same for TiVNbTa, TiVNbMoTa, TiVNbZr, and TiVNbMo: Arc melted from the mixture. "Powder samples were ground from the as-cast ingots". I will use the "as-cast" condition, as that is the intrinsic state of the alloy that the properties are associated with.
**Final Data Points and JSON Output**
I've carefully examined all the pertinent text, ensuring I understand the preparation methods, the alloy's conditions for property measurements, and the meaning of each parameter. After all that analysis, I'm ready to craft the JSON objects, and I've decided to exclude MoNbTaW from the list of *primary alloys analyzed in this paper*.
Here's my final list:
**1. TiVNbTa**
```json
{
  "name": "TiVNbTa",
  "nominal_composition": "Ti1.0V1.0Nb1.0Ta1.0",
  "measured_composition": "Not found",
  "lattice_constant": "3.221 |\AA{}|",
  "phases": "bcc solid-solution structure",
  "alloy_condition": "As-Cast",
  "doi": "10.1016/j.jallcom.2020.155970"
}
```
```

```
**2. TiVNbMoTa**
```json
{
  "name": "TiVNbMoTa",
  "nominal_composition": "Ti1.0V1.0Nb1.0Mo1.0Ta1.0",
  "measured_composition": "Not found",
  "lattice_constant": "3.197 |\AA{}|",
  "phases": "bcc solid-solution structure",
  "alloy_condition": "As-Cast",
  "doi": "10.1016/j.jallcom.2020.155970"
}
```
3. NiCoFeCrMn
```json
{
  "name": "NiCoFeCrMn",
  "
```

Figure S11: The extended reasoning history from Gemini 2.5 Flash with dynamic thinking budget for the Zhang et al. [9] publication showing its internal reasoning process before generating the final extraction output..
```

S6. EXTENDED REASONING USING CLAUDE 4.5 SONNET MODEL

```
Let me carefully read through the PDF to extract information about all high-entropy alloys discussed in the results section.
From the Results and Discussion section, I can identify the following alloys:
1. **TiVNbTa** - mentioned in Fig. 1a and b, neutron diffraction profiles
   - Lattice constant at room temperature: 3.2206 (9) Å
   - Average volume: 16.70 (2) Å$^3$/atom
   - Structure: bcc
   - Nominal composition: Ti1.0V1.0Nb1.0Ta1.0
2. **TiVNbMoTa** - mentioned alongside TiVNbTa
   - Structure: bcc
   - Nominal composition: Ti1.0V1.0Nb1.0Mo1.0Ta1.0
3. **NiCoFeCrMn** - mentioned as a comparison fcc alloy
   - Debye-Waller factor at room temperature: 0.40 Å$^3$
   - Structure: fcc
   - Nominal composition: Ni1.0Co1.0Fe1.0Cr1.0Mn1.0
4. **TiVNbZr** - mentioned in the text
   - Structure: bcc
   - Nominal composition: Ti1.0V1.0Nb1.0Zr1.0
5. **TiVNbMo** - mentioned in the text
   - Structure: bcc
   - Nominal composition: Ti1.0V1.0Nb1.0Mo1.0
All alloys appear to be from arc-melted/as-cast condition based on the experimental details.
DOI: 10.1016/j.jallcom.2020.155970
Let me extract specific information for each alloy from the results section.
```

Figure S12: The extended reasoning history from Claude 4.5 Sonnet with 1024 thinking budget for the Zhang et al. [9] publication showing its internal reasoning process before generating the final extraction output.

S7. POTENTIAL IMPROVEMENTS

We identify four potential strategies for improving data extraction reliability.

Strategy 1: Expand Training Set Diversity. The optimization process would benefit from expert-labeled papers representing greater diversity in: (*i*) authorship, (*ii*) target applications and material systems, (*iii*) journal formatting conventions, and (*iv*) data presentation styles (text vs. tables vs. figures). Increased diversity would provide the prompt optimization algorithm with broader exposure to edge cases, ambiguous phrasings, and varied reporting conventions, improving generalization to the full literature collection. Zheng et al. [10] made a similar recommendation in their study.

Strategy 2: Implement Output-Level "Gradient" Optimization. We can extend the TextGrad optimization framework to include a Module 4 that uses output-level "gradients" in addition to prompt-level optimization. As demonstrated in Section 4, the `EvaluatePdfOutputLoss` function successfully identifies many extraction errors in generated outputs. Implementing an additional optimization loop that iteratively refines outputs based on these loss signals could enable automatic error correction. However, this approach has two limitations: (*i*) when the LLM rationalizes its errors (where the model asserted that 0.319 Å was correct), the loss function fails to generate corrective "textual gradients", maintaining the error, and (*ii*) additional optimization is costly as each refinement requires multiple API calls to commercial LLM services.

Strategy 3: Prioritize Original Publisher PDFs Over XML Reconstructions. We identify that there is value in using publisher-provided PDFs rather than PDFs reconstructed from XML sources. While XML-to-PDF reconstruction successfully preserved text and tabular data for most papers, formatting fidelity was inconsistent. This was especially true for publications with equations, special symbols, and complex multi-column tables. These formatting artifacts can potentially impair LLM extraction performance. A representative example from Cui et al. [8] illustrates this impact. Extraction from the XML-reconstructed PDF identified only three of six HEA compositions (recall = 0.50), with the loss function (Figure S10 in SI) confirming the missed entries. In contrast, the same extraction pipeline applied to the publisher's original PDF correctly identified all six compositions in a single pass, indicating that formatting quality directly affects extraction reliability. The exact mechanism that

contributes to this improvement needs further investigation.

Strategy 4: Leverage Advances in LLM Model Capabilities. LLM models continue to improve rapidly, with each generation showing enhanced reasoning, improved adherence to prompt specifications, and multimodal understanding [11, 12]. Future LLM models may substantially improve extraction performance through better contextual understanding, reduced hallucination rates, and improved handling of complex scientific content. Our cross-model evaluation (discussed in Testing Prompt Transferability subsection in the main manuscript) demonstrates this trend, with newer models achieving superior performance even with unoptimized prompts. In addition, retrieval-augmented generation (RAG) offers another promising approach to reduce semantic hallucination by ensuring LLM outputs are based on actual document content rather than model memory [13]. Nonetheless, the high cost of commercial LLM APIs could restrict their broad use for large-scale data extraction in resource-constrained research environments. Fine-tuning open-source LLMs presents a promising and cost-effective alternative that deserves further exploration.

S8. PERFORMANCE OF THE TRAINED ML MODELS IN PREDICTING LATTICE CONSTANTS

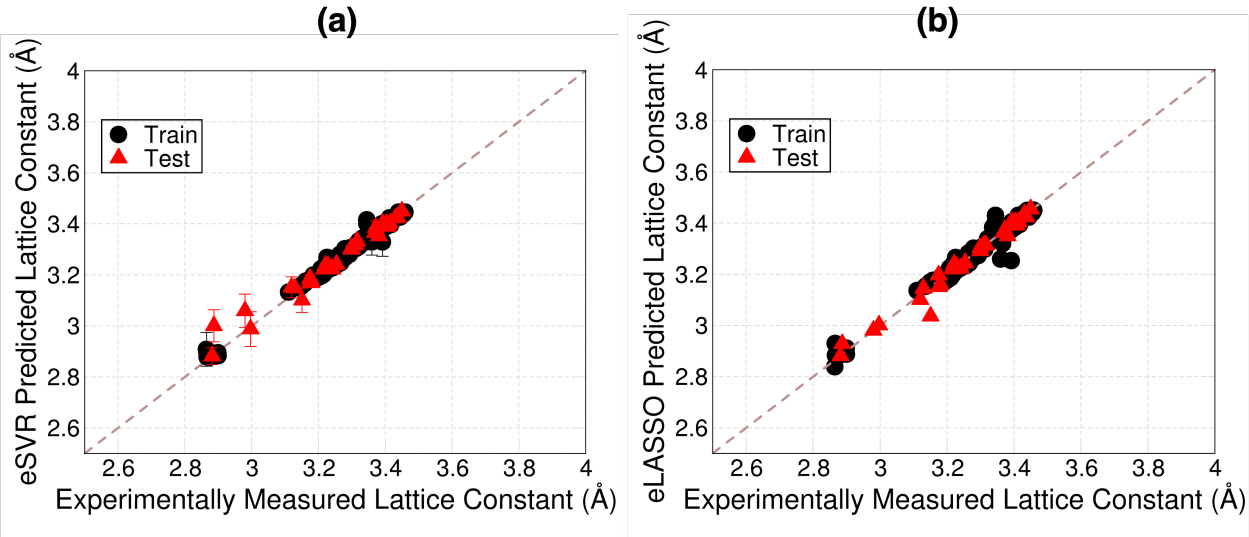

Figure S13: Parity plot comparing (a) eSVR-predicted (y-axis) and (b) eLASSO-predicted lattice constants (y-axis) against experimental lattice constants (x-axis) for LLM-extracted as-cast single-phase BCC HEAs. Data points falling on the diagonal dashed line (y = x) represents perfect prediction. Filled black circles denote training set data (127 data points), and filled red triangles indicate test set data (32 data points). Error bars represent prediction uncertainty quantified as the standard deviation across the 100 ensemble SVR models in (a), and 1000 ensemble LASSO models in (b). Strong clustering along the diagonal for both training and test sets demonstrates good model generalization (test $R^2 = 0.969$ and 0.974 for eSVR and eLASSO, respectively).

S9. SUGGESTED STEPS PRIOR TO PERFORMING DOWNSTREAM TASKS

Based on the lessons learned from our detailed analysis of the as-cast single-phase BCC HEAs, we recommend the following steps prior to performing any downstream tasks with the extracted lattice constants.

- Randomly select n publications (where the exact value of n depends on the budget) and verify that the LLM has accurately extracted the lattice constants from those publications. For convenience, the DOI is provided in the same row as the corresponding lattice constant data. Ideally, both precision and recall should be close to 1.

- If the budget allows, conduct the `ExtractPDFOutputLoss` analysis for those n publications to evaluate the accuracy of the extracted outputs.

- Use the Check_Composition_Consistency.py provided with the source code to verify compositional information (the GitHub link is given in the Code Availability section in the main manuscript). Manually check if the L1 distance (D_{L1} in Equation 4 in the main manuscript) or cosine similarity (CosSim in Equation 5 in the main manuscript) are unusual. This utility python script has provided useful parsing strategies that normalizes composition and will flag rows with inconsistent compositions.

S10. PSEUDOCODE FOR THE ITERATIVE PROMPT OPTIMIZATION

The pseudocode is shown in Figure S14.

Iterative Prompt Optimization

Inputs: train_loader
prompt ← initial_prompt
model ← LLM(prompt)
optimizer ← TextualGradientDescent({prompt})
loss_fn ← EvaluatePdfExtracationLoss()

1: **for** epoch ← 1 **to** E **do**
2: **for each** $(\text{batch}_x, \text{batch}_y)$ **in** train_loader **do**
3: $\mathcal{L} \leftarrow [\,]$
4: **for each** (x, y) **in** $(\text{batch}_x, \text{batch}_y)$ **do**
5: $\hat{y} \leftarrow \text{model}(x)$
6: $\ell \leftarrow \text{loss\_fn}(\text{prompt}, x, \hat{y}, y)$
7: $\text{append}(\mathcal{L}, \ell)$
8: **end for**
9: $L \leftarrow \sum_{\ell \in \mathcal{L}} \ell$
10: Backpropagate(L)
11: UpdateParameters(optimizer)
12: **end for**
13: **end for**

Figure S14: Pseudocode for iterative prompt optimization (Module 2) based on the TextGrad framework [14]. The inputs are: x (original PDF), y (expert-labeled ground truth data), and $\hat{y}$ (LLM-generated output). The training data are loaded in batches via `train_loader`, which provides batches of PDFs (batch_x) and corresponding expert labels batch_y. For each (x, y) pair in a batch, the PDF x is processed by the LLM using the current prompt to generate $\hat{y}$, and the `EvaluatePdfExtractionLoss` function computes the loss ℓ by comparing $\hat{y}$ to the expert label y. Individual losses within the batch are accumulated in list $\mathcal{L}$ and summed to produce batch loss L (line 9). The loss is then backpropagated through the prompt using textual gradient descent, and the prompt parameters are updated by the optimizer. This process repeats for E epochs across all training batches. For implementation details, see Yuksekgonul et al. [14].

S11. PROMPTS USED FOR THE LOSS FUNCTION DURING PROMPT OPTIMIZATION

```
Below is the output from a data extraction task for a given PDF file, the PDF data, and the result from human of this task.
            You need to evaluate the output and the result from human.
          ONLY compare presented entries, do not perform task yourself.
Is the output matches with human results? i.e., Does the reported parameter match with the PDF data and the human?
i.e., does the number of material matches with PDF and human annotation? Does the lattice constant matches with the PDF and human annotation?
              Say match if it does, otherwise say it doesn't match.
                                  Be super consice.
```

Figure S15: Prompts for the loss function during prompt optimization (Module 2).

```
Below is the output from a data extraction task for a given PDF file, and the PDF itself.
You need to evaluate the output according to the provided PDF for each extracted material, be super concise:
 Does the output matches the format provided in the question? Answer yes or no.
        Are the extract material high-entropy alloys? Answer yes or no.
Are the alloy composition in the output correct and reflect the material? Answer yes or no.
Are the lattice constants value in the output formatted in angstrom? Answer yes or no.
Are the lattice constants in the output truly the lattice constant of the material not others? Answer yes or no.
Are the phase information in the output correct and reflect the material? Answer yes or no.
Are the alloy condition in the output correct and reflect the material? Answer yes or no.
   Is the DOI in the output matches with the original paper? Answer yes or no.
  Is there any HEA high-entropy alloys missed in the provided ouput? yes or no.
```

Figure S16: Prompts for the loss function for output evaluation.

S12. DEFINITION OF CONFUSION MATRIX FOR RETRIEVAL EVALUATION

The confusion matrix, which defines true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), is presented in Table S3 below [15, 16].

TABLE S3: Definition of various terms in the confusion matrix for information retrieval evaluation.

	Relevant	**Non-relevant**
Retrieved	True Positives (TP)	False Positives (FP)
Not Retrieved	False Negatives (FN)	True Negatives (TN)

S13. DESCRIPTOR SELECTION FOR THE DOWNSTREAM MACHINE LEARNING TASK

The descriptor selection pipeline employs Recursive Feature Elimination with Cross-Validation (REFCV), as implemented in `scikit-learn`[17]. This module selects descriptors by iteratively removing the least important ones through a recursive process. This method is commonly applied in various machine learning tasks in the literature [18–20]. In our study, we utilize REFCV in conjunction with `RandomForestRegression` [21], employing a three-fold cross-validation strategy (CV=3). We started with 36 descriptors and the REFCV approach downselected six descriptors for training the ML models.

REFERENCES

[1] Allen Institute for AI. Asta: Agentic AI research assistant. `https://asta.allen.ai`, 2025. Accessed 4 October 2025.

[2] Mujin Yang, Chao Huang, Jiajia Han, Haichen Wu, Yilu Zhao, Tao Yang, Shenbao Jin, Chenglei Wang, Zhou Li, Ruiying Shu, Cuiping Wang, Huanming Lu, Gang Sha, and Xingjun Liu. Development of the high-strength ductile ferritic alloys via regulating the intragranular and grain boundary precipitation of G-phase. *Journal of Materials Science & Technology*, 136:180–199, 2023.

[3] S. Pan, J.Q. Zhang, B.B. He, and M.X. Huang. High strain gradient induced nanograin in a cocrnivc medium entropy alloy with lamellar carbide. *Materials Science and Engineering: A*, 852:143692, 2022.

[4] C.L. Wu, S. Zhang, C.H. Zhang, H. Zhang, and S.Y. Dong. Phase evolution and cavitation erosion-corrosion behavior of fecocralniti$_t$*extrmx* high entropy alloy coatings on 304 stainless steel by laser surface alloying. *Journal of Alloys and Compounds*, 698:761–770, 2017.

[5] Jongun Moon, Elena Tabachnikova, Sergii Shumilin, Tetiana Hryhorova, Yuri Estrin, Jamieson Brechtl, Peter K. Liaw, Wenqing Wang, Karin A. Dahmen, Alireza Zargaran, Jae Wung Bae, Hyeon-Seok Do, Byeong-Joo Lee, and Hyoung Seop Kim. Deformation behavior of a co-cr-fe-ni-mo medium-entropy alloy at extremely low temperatures. *Materials Today*, 50:55–68, 2021.

[6] David Piorunek, Jan Frenzel, Niels Jöns, Christoph Somsen, and Gunther Eggeler. Chemical complexity, microstructure and martensitic transformation in high entropy shape memory alloys. *Intermetallics*, 122:106792, 2020.

[7] Yiling Jiang, Haibo Tang, Zhuo Li, Dongdong Zheng, and Yuxuan Tu. Additive manufactured Mg-Gd-Y-Zr alloys: Effects of Gd content on microstructure evolution and mechanical properties. *Additive Manufacturing*, 59:103136, 2022.

[8] Peng Cui, Yimo Ma, Lijun Zhang, Mengdi Zhang, Jiantao Fan, Wanqing Dong, Pengfei Yu, and Gong Li. Microstructure and mechanical behaviors of $CoFeNiMnTi_xAl_{1-x}$ high entropy alloys. *Materials Science and Engineering: A*, 731:124–130, 2018.

[9] F.X. Zhang, Y. Tong, M. Kirkham, A. Huq, H. Bei, W.J. Weber, and Y. Zhang. Structural disorder, phase stability and compressibility of refractory body-centered cubic solid-solution alloys. *Journal of Alloys and Compounds*, 847:155970, 2020.

[10] Zhiling Zheng, Oufan Zhang, Christian Borgs, Jennifer T Chayes, and Omar M Yaghi. ChatGPT chemistry assistant for text mining and the prediction of MOF synthesis. *Journal of the American Chemical Society*, 145(32):18048–18062, 2023.

[11] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[12] Mian Zhang, Shujian Liu, Sixun Dong, Ming Yin, Yebowen Hu, Xun Wang, Steven Ma, Song Wang, Sathish Reddy Indurthi, Haoyun Deng, et al. Complex logical instruction generation. *arXiv preprint arXiv:2508.09125*, 2025.

[13] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[14] Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Pan Lu, Zhi Huang, Carlos Guestrin, and James Zou. Optimizing generative AI by backpropagating language model feedback. *Nature*, 639(8055):609–616, 2025.

[15] Xin Wang, Anshu Raj, Matthew Luebbe, Haiming Wen, Shuozhi Xu, and Kun Lu. Reliable End-to-End Material Information Extraction from the Literature with Source-Tracked Multi-Stage Large Language Models, 2025.

[16] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[17] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[18] Junya Wang, Pengcheng Xu, Xiaobo Ji, Minjie Li, and Wencong Lu. Feature Selection in Machine Learning for Perovskite Materials Design and Discovery. *Materials*, 16, 2023.

[19] Sofia Sheikh, Brent Vela, Vahid Attari, Xueqin Huang, P. Morcos, James Hanagan, Cafer Acemi, Ibrahim Karaman, A. Elwany, and Raymundo Arroyave´. Exploring chemistry and additive manufacturing design spaces: a perspective on computationally-guided design of printable alloys. *Materials Research Letters*, 12:235 – 263, 2024.

[20] Elliot H. E. Farrar and Matthew N. Grayson. Machine learning and semi-empirical calculations: a synergistic approach to rapid, accurate, and mechanism-based reaction barrier prediction. *Chemical Science*, 13:7594 – 7603, 2022.

[21] Kevin Maik Jablonka, Daniele Ongari, Seyed Mohamad Moosavi, and Berend Smit. Big-Data Science in Porous Materials: Materials Genomics and Machine Learning. *Chemical Reviews*, 120:8066 – 8129, 2020.