arXiv:2512.22130v1 [cs.DL] 5 Dec 2025

基于专家知识的自动提示工程：使用大语言模型从科学出版物中提取高熵合金的晶格常数

Shunshun Liu[1], TalonR. Booth[1], YangfengJi[2],Wesley Reinhart[3,4], PrasannaV.Balachandran[1*]1*材料科学与工程系，弗吉尼亚大学，夏洛茨维尔，22903，弗吉尼亚州，美国。[2]计算机科学系，弗吉尼亚大学，夏洛茨维尔，22903，弗吉尼亚州，美国。[3]材料科学与工程系，宾夕法尼亚州立大学，斯泰特学院，16802，宾夕法尼亚州，美国。[4]计算与数据科学研究所，宾夕法尼亚州立大学，斯泰特学院，16802，宾夕法尼亚州，美国。

通讯作者。电子邮箱：pvb5e@virginia.edu；

摘要

大语言模型（LLM）在从出版物中提取科学数据方面展现出应用前景，但仍依赖人工优化提示词。我们提出了一种以专家知识为基础的自动提示优化框架，以提高大语言模型实体抽取的可靠性。我们以高熵合金晶格常数提取为测试案例，通过对七篇经专家标注的出版物进行反馈迭代，优化了 Claude 3.5 Sonnet 的提示词。尽管优化预算有限，召回率仍从 0.27 提高到 > 0.9，表明小规模的专家整理数据集也能带来显著提升。随后，我们将该方法用于从 2,267 篇出版物中提取晶格常数，获得了 1,861 种成分的数据。优化后的提示词也能有效迁移至更新的模型：Claude 4.5 Son-net、GPT-5 和 Gemini 2.5 Flash。分析发现了大语言模型的三类错误：上下文幻觉、语义误解和单位换算错误，这凸显了制定验证方案的必要性。这些结果表明，反馈引导的提示优化是一种低成本、可迁移的方法

，可用于可靠的科学数据提取，并为复杂的大语言模型辅助研究任务提供可扩展的途径。

关键词： 大语言模型、提示优化、高熵合金、晶格常数、科学数据提取

引言

基于实验数据的材料信息学研究历来依赖于从科学出版物中手动提取数据。这一过程既繁琐又耗时。因此，将这种方法扩展到从数千篇出版物中提取数据仍不切实际。在过去十年中，研究人员一直致力于实现从非结构化来源（如期刊出版物）中自动提取数据。最初，人们开发了依赖专家标注的自然语言处理（NLP）技术 [1–5], 近年来，研究重点转向利用大语言模型（LLM）来完成这一任务 [6–18]。我们对大语言模型在材料科学文献中的应用进行了调查（补充信息（SI）中的图 S1），发现自动化知识提取和数据提取是最普遍的应用场景。这些研究表明，从科学出版物中进行基于LLM的数据提取可以达到合理的准确率，但其性能在很大程度上取决于提示词的质量。提示词是提供给大语言模型的一组结构化指令，用于规定任务、定义约束条件，并说明输出格式和内容 [19, 20]。有效的提示词通常包含四个要素 [6, 10, 19–21]：**(1)** 指令，用于定义任务和期望的行为；**(2)** 上下文，用于提供相关背景信息；**(3)** 输入数据，用于处理；以及 **(4)**输出指标，用于规定响应的期望格式和结构。

我们的文献调查发现，近期有多篇出版物提出了从科学出版物中提取材料科学数据的提示工程系统方法。下面我们将详细讨论两项具有代表性的研究。Zheng 等人 [6]针对化学领域的应用，基于三项核心原则开发了人在回路提示工程策略：(*i*) 精心设计提示词，避免提出可能产生虚构内容的问题，从而尽量减少幻觉；(*ii*) 提供详尽的指令，说明上下文和期望的答复格式；以及 (*iii*) 使用定义明确的模板规定结构化输出要求，以便于数据提取。他们的工作流程首先确定出版物中的相关章节（即包含金属有机框架（MOF）合成条件的章节），确认其中存在数据，然后从这些章节中提取金属有机框架合成参数。通过结合使用 GPT-3.5 和 GPT-4，以及经人工反馈增强的交互式提示词优化策略，他们展现出在实现稳健性能指标方面的显著成效，并使用精确率、召回率和 F1 分数对这些指标进行了讨论。Polak 和 Morgan [10] 开发了一种两阶段零样本提示工程策略。第一阶段使用相关性提示识别包含目标数据的句子，第二阶段则使用问答提示提取具体信息。他们通过构建两个数据库展示了这一方法：一个收录金属玻璃的临界冷却速率，另一个收录高熵合金

的屈服强度。研究人员使用 GPT-3.5（gpt-3.5-turbo-0301）、GPT-4（gpt-4-0314）和 LLaMA2-chat（70B），并通过精确率和召回率指标评估了提取准确度。Ha 等人 [22]（纳米毒性）、Zhang 等人 [23] （磁性材料）、Liu 等人 [13]（金属玻璃）、Gupta 等人 [24] （聚合物）以及 Prasad 等人 [25]（钛基合金）的其他研究进一步证明，精心构建的提示词有助于从科学文本中大规模提取数据。

当前的提示工程方法尽管取得了进展，但仍存在一项重大局限：它们需要大量人工参与来改进提示词，而且通常是针对单篇出版物进行优化，而非面向一组出版物进行优化。为解决这一局限，我们提出一种提示工程策略，将领域专家构建的数据融入自动优化流程，而不是依赖事后的人机协同改进。我们的工作流程如图1所示。在 TextGrad 框架 [26], 的基础上，该框架通过对大语言模型生成的“文本梯度”进行反向传播来实现提示优化。我们的工作流程利用一个经过精心整理的小型专家依据数据集，该数据集取自多篇科学出版物，可在无需人工参与的情况下自动、迭代地改进提示词。随后，这些优化后的提示词可用于从数千篇科学出版物中大规模自动化数据提取。我们以高熵合金作为测试案例来展示这一方法。高熵合金是一类新兴材料，具有重要的技术应用价值 [27–29]。在提示词中，我们使用“高熵合金”一词，大语言模型会将其理解为涵盖多主元合金（MPEAs）和成分复杂合金（CCAs），因为这些术语在文献中可以互换使用。为保持一致，本文始终使用高熵合金这一称谓，同时认为其广义上涵盖多主元合金和/或成分复杂合金。

我们的目标是开发一种以专家知识为基础的自动提示优化框架，利用大语言模型从数千篇高熵合金出版物中提取实验测得的室温晶格常数数据。现有的高熵合金数据库，例如 Borg 等人建立的数据库 [30], （包含力学性能和相信息），以及 Polak 和 Morgan 建立的数据库 [10], （包含屈服强度数据和带注释的源文段），均未提供实验合成合金的晶格常数数据。要推动合金开发，建立一个包含实验高熵合金文献中晶格常数数据的高熵合金数据集至关重要[31]。这些数据是利用机理模型预测固溶强化效应的重要输入 [32, 33] ，也有助于估算沉淀硬化体系中基体与析出相之间的共格应变关系 [34]。然而，人工从迅速增长的高熵合金科学出版物中提取晶格常数数据并不现实，这已成为材料设计与发现工作的瓶颈。

我们的专家指导的提示优化框架由三个模块组成，如图1所示。更多详情请参阅方法部分。模块1（专家标注）：领域专家通过调查少量科学出版物，手动创建提示训练数据集。此步骤确定待提取的命名实体、所需的数据字段，以及

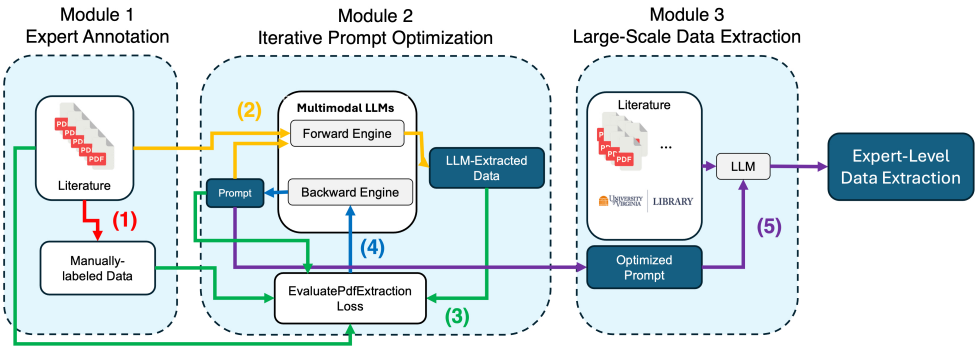


图 1 专家指导的提示优化与大规模数据提取的示意工作流程。括号中的数字表示工作流程的顺序步骤，箭头颜色与之对应。各模块的详细信息见正文。

指导后续提示优化的结构化格式。模块 2（迭代式提示优化）：重点是利用大语言模型和 TextGrad 框架 [26]进行迭代式提示优化。此模块使用与模块 1 中由领域专家标注的相同出版物。它需要两个输入：(*i*)用于构建模块 1 专家标注数据集的 PDF 格式科学出版物，以及 (*ii*)初始提示词。通过多模态大语言模型系统以及利用前向引擎和反向引擎的工作流程，初始提示词得到迭代式自动优化。Claude 3.5 Sonnet [35] 既作为数据提取的前向引擎，也作为生成改进建议的反向引擎。使用TextGrad 损失模块实现的`EvaluatePdfExtractionLoss`函数会将大语言模型提取的数据与专家标注的真实标注进行比较，以计算“文本梯度”。这一优化循环会持续进行，直到迭代优化后的提示词达到令人满意的提取性能。模块 3（大规模数据提取）：将模块 2 优化后的提示词用于从数千篇科学出版物中提取数据。这种依次执行的三模块方法，能够利用有限的专家整理数据系统性地优化提示词，随后在数千篇科学文献中进行大规模提取。尽管本文使用 Claude 3.5 Sonnet 展示了这一方法，但商业或开源大语言模型的任意组合也都可以用作前向引擎或反向引擎。

将这一方法应用于 2,267 篇科学出版物后，我们提取了 1,861 种高熵合金成分的晶格常数数据。除晶格常数外，我们还收集了化学成分（包括名义成分和实测成分，如有）、相信息及加工技术。按照文献中的成熟方法，我们使用精确率、召回率和 F1 分数指标评估了提取性能。对这 1,861 条记录进行分析后，我们确定了 186 种在铸态加工条件下呈现单相体心立方（体心立方）结构的材料。为展示提取数据集在下游应用中的效用，我们构建了数据驱动的机器学习模型，以铸态单相体心立方高熵合金成分为输入来预测晶格常数；这些成分是高温结构应用的潜在候选材料 [36–38]。

尽管 Claude 3.5 Sonnet 大语言模型在大规模数据提取方面表现出色，我们也讨论了三个可能危及抽取可靠性的挑战：（i）上下文幻觉，即模型生成了在物理上看似合理、但缺乏文本依据的数值 [39, 40]；（ii）语义误解 [39, 40], ，即模型误解了技术术语或隐含上下文；以及（iii）单位不一致，给下游处理带来了困难。深入分析这些局限性，有助于了解当前基于大语言模型从科学文献中提取数据所面临的挑战。

最后，尽管更新一代的大语言模型展现出强大的零样本推理能力，但某些科学任务仍然颇具挑战，例如综合跨模态信息 [17,18]。因此，要推动模型性能达到实际应用的上限，系统化、自动化的提示优化流程至关重要。本文实现了这样一套流程，并以高熵合金晶格常数提取作为代表性基准测试对其进行了验证。尽管提示优化工作流程基于 Claude 3.5 Sonnet 模型，但在当前的大语言模型上测试时，优化后的提示词展现出出色的迁移能力，包括 Claude 4.5 Sonnet、Gemini 2.5 Flash 和 GPT-5。我们的结果为领域科学家提供了前进方向，以应对前沿模型能力的快速提升。

结果

初始提示与优化后的提示词

初始提示和优化后的提示词分别见图2和图3。初始提示由作者依据领域专业知识构建；设计过程中既未咨询大语言模型，也未采用系统化的提示工程原则。对初步提取尝试的分析发现了三个会降低输出质量的缺陷：（i）指令含糊且提取规则不完整，（ii）输出格式不一致，以及（iii）提取范围界定不清。下面将讨论这些局限，以及优化后的提示词中纳入的相应改进。

```
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"
}
```

图2 完整的初始提示。补充信息中的图S2提供了机器可读版本。

初始提示词产生歧义的原因之一，是未定义技术术语以及缺少处理规则。例如，“单相”这一短语没有得到明确界定，对于具有多个已报告相的材料，也没有提供清晰的指令。优化后的提示词解决了这一问题

(a)

```
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.
```

(b)

```
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
        }
    },
    ...
]
```

(c)

```
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
        }
    }
]
```

(d)

```
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
        }
    }
]
```

图 3 完整优化后的提示词的精简版和简化可视化图，分为四个面板：(a) 任务指令、提取范围以及数据提取方案的定义。(b) JSON 格式的提取数据示例。(c) 补充指令以及演示正确输出的 JSON 示例。(d) 演示如何处理缺失值的 JSON 示例。补充信息中的图 S3 提供了机器可读版本。

，将范围扩展至所有高熵合金（图3，第1行），同时移除了含义模糊的“单相”限制。这一调整是合理的，因为单相高熵合金属于复杂高熵合金相空间的子集，可以根据提取的相信息在后处理中进行筛选。此外，初始提示词缺少处理缺失数据或每篇出版物包含多个条目的规则。优化后的提示词通过文本指令（第11行、第39行）和具体的格式示例明确处理了这些情况：第15—37行展示了多个材料条目的处理方式，而第63—83行说明了正确的缺失值表示法。

第二个弱点涉及输出格式。初始提示定义了数据模式，但未规定各个值的格式。研究发现，缺乏明确指导时，大语言模型容易生成不一致的输出。

优化后的提示词通过详细的格式要求（第 3–7 行）消除了这一弱点，确保输出统一并符合预期的数据模式。

第三个弱点是提取范围未明确定义。如果没有明确的边界，大语言模型可能会从参考文献列表或引文中提取无关数据，收录论文中提及但未进行实验研究的材料。为解决这一弱点，优化后的提示词明确了范围优先级（第11、39行）：“从结果中提取信息，优先关注文本，其次是表格，最后是图。”

专家标注训练集上的数据检索性能

我们在专家标注训练集上评估了初始提示词和优化后的提示词的数据提取性能。该训练集包含七篇研究论文，其中有22种单相高熵合金。使用初始提示词的 Claude 3.5 Sonnet 模型仅从七篇出版物中提取了七条记录。初始提示词的提取指标见表1，结果显示其召回率存在严重局限：专家识别出的高熵合金条目中，只有0.273（27.3%）被成功提取。这意味着约73%的相关数据未被发现，凸显了系统化提示工程的必要性。召回率较低的原因之一是初始提示词中的提取规则不完整。具体而言，尽管许多论文报告了多种高熵合金成分，该提示词却隐含地将提取范围限制为每篇出版物中的一种高熵合金成分。提取出的七条记录中，有一种高熵合金成分为AlHfNbTaTiZr，该成分取自 Lin 等人的研究。 [41]。我们有意将这种高熵合金成分排除在专家构建的训练集之外，因为作者报告了两个不同的体心立方晶格常数，这表明其具有多相微观结构，而非单相体心立方结构。我们将此视为该成分为两相体心立方结构、而非单相的证据。初始提示词共提取七条记录，其中一条提取错误，因此其总体精确率为0.857。

相比之下，优化后的提示词从这七篇出版物中提取了 23 条单相高熵合金成分记录，超过了专家标注的 22 种单相高熵合金成分。提取的 23 条记录中，有两种高熵合金虽在论文中被提及，但并未在相关研究中进行实验表征，因此属于假正例（FP）。此外，优化后的提示词还提取了 Lin 等人 [41]的 AlHfNbTaTiZr，并将其归类为单相体心立方，尽管此前讨论过两种不同的晶格常数。

根据特定的单相高熵合金标准进行评估时，优化后的提示词精确率为 0.870（20/23），召回率为 0.909（20/22）。然而，优化后的提示词中“提取结果部分讨论的所有高熵合金”这一指令，将实际范围扩展到了单相成分之外。若针对所有高熵合金记录进行评估（不论是单相还是多相），优化后的提示词从总计 37 种成分中提取了 36 种，且只有两个假正例，精确率为 0.944，召回率为 0.919。这种更广泛的评估更能反映优化后的提示词按照其指令设定时的实际表现。

实体级性能（表1）显示，名义成分和工艺条件均保持了完美的精确率（1.000），而晶格常数的精确率

略有下降（0.912）。进一步分析后，我们发现，精确率降低是由多相处理相关问题造成的。例如，Zuo 等人 [42] 研究了一系列 $Al_xCoFeNi$ 高熵合金成分。他们发现，当 x = 0.5 且为 0.75 时，这些合金呈现由面心立方相和体心立方相组成的双相微观结构（Zuo 等人 [42]中的表 1）。此外，在 Zuo 等人 [42]的表 2 中，作者报告了与 x = 0.5 且为 0.75 对应的面心立方相和体心立方相各自不同的晶格常数。然而，优化后的提示词未能准确地将每个晶格参数与其对应的相关联起来，导致晶格常数提取错误。因此，表 1中晶格常数和相的精确率略低。

表1 使用 Claude 3.5 Sonnet 模型比较初始提示与优化后的提示词在实体层面的提取性能。

命名实体	初始提示			优化后的提示词		
	精确率	召回率	F1 分数	精确率	召回率	F1 分数
名义成分	1.0	0.273	0.429	1.0	0.919	0.958
晶格常数	1.0	0.273	0.429	0.912	0.838	0.873
相	1.0	0.273	0.429	0.971	0.892	0.930
合金加工条件	1.0	0.273	0.429	1.0	0.919	0.958

大规模数据提取

在通过数据提取工作流程的模块 3（图1）处理的 2,267 篇出版物中，有 142 篇被 Claude 3.5 Sonnet 模型拒绝。这些出版物之所以被拒绝，是因为它们要么是超出上下文长度限制的综述文章，要么在大语言模型进行的评估中被归类为与高熵合金无关。大规模数据提取模块从 2,125 篇出版物中共发现 4,648 条合金条目。其中，2,787 条（60.0%）不包含晶格常数数据，而 1,861 条（40.0%）包含晶格常数，如图4（a）所示。在包含晶格常数数据的 1,861 条条目中，有 408 条属于面心立方（FCC）合金，311 条属于体心立方合金，1,146 条属于其他相（包括多相混合物和其他有序结构类型）。如图4（b）所示。在 311 条单相体心立方合金条目中，186 条为铸态，111 条通过其他方法（如粉末加工、选择性激光熔化、退火）加工或后处理。此外，14 条条目未报告任何加工信息，如图4（c）所示。

图 4（d）展示了铸态单相体心立方合金的晶格常数分布。大多数数值位于 2.8 至 3.5 埃之间（$^{\circ}$A），符合预期的体心立方晶格参数范围。然而，有 21 条记录的晶格常数接近 0。3 $^{\circ}$A，这表明 Claude 3.5 Sonnet 模型未能将提取过程中以纳米为单位报告的晶格常数换算为 $^{\circ}$A。我们共发现 327 条记录的晶格常数在物理上不合理：其中 17 条记录（来自

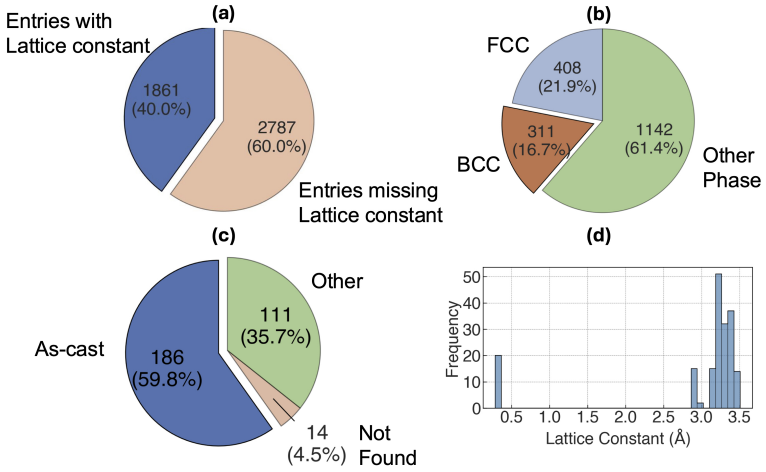


图 4 从 2,267 篇出版物中提取的高熵合金数据概览。(a) 饼图显示了包含和不包含晶格常数数据的条目数量占比：在提取的 4,648 个成分中，1,861 个（40.0%）包含晶格常数数据，2,787 个（60.0%）缺少此类数据。(b) 饼图显示了 1,865 个包含晶格常数的条目的相分布：311 个为单相体心立方（16.7%），408 个为单相面心立方（21.9%），1,142 个为其他相（61.4%），包括多相混合物、有序结构和非晶结构。(c) 饼图显示了 311 种单相体心立方合金的加工条件分布：186 种为铸态（59.8%），111 种采用其他方法加工（如选择性激光熔化和粉末加工），或进行了后处理热处理（如退火）（35.7%）；另有 14 种未报告加工信息（4.5%）。(d) 直方图显示了 186 种铸态单相体心立方合金的晶格常数分布。对于体心立方结构，大多数数值落在 2.8–3.5 $^{\circ}$A 范围内。有 21 个条目的数值接近 0.3 $^{\circ}$A，这代表一种系统性的单位换算错误，即大语言模型未能将纳米数值换算为 $^{\circ}$A 单位。

七篇出版物）的数值为 >10 $^{\circ}$A，另有 310 条记录（来自 166 篇出版物）的数值为 <1 $^{\circ}$A。我们人工检查了所有 7 篇数值异常偏大的出版物，以及随机抽取的 10 篇数值异常偏小的论文，发现大多数离群值源于原始文献中以纳米或皮米为单位报告的晶格常数。尽管提示词中明确要求以$^{\circ}$A 为单位报告数值，Claude 3.5 Sonnet 模型仍未能执行必要的单位换算，构成一种系统性的格式错误。除了单位换算错误外，人工验证还发现了另外两类代表大语言模型提取挑战的系统性问题：（1）上下文幻觉；（2）语义误解。这些挑战及其代表性示例总结于表 2。

为评估我们的数据提取性能，我们随机抽取了46篇不同的出版物，重点关注铸态单相体心立方合金条目，这是下游任务的关注重点。在此次评估中，我们排除了晶格常数异常的异常值条目。在抽取的46篇出版物中，我们人工识别出44种铸态单相体心立方高熵合金成分。我们没有将相和合金加工条件视为两个独立实体，而是将二者合并为一个复合判据，称为“铸态单相体心立方”。这一简化

表 2 本研究中遇到的三项大语言模型科学数据提取挑战总结。

Type	说明	具体示例
上下文幻觉现象	The 模型分配 一个正确的——看起来 数字到 the 错误的实体 由于链接错误或错误地——归属。	The LLM 错误地 被解-释 the 晶格 常数 的 镍$_{16}$钛$_{6}$补充信息$_{7}$ 析出相 相 as that 的 host 基体 ，其具有不同的成分。（Yang 等人 [43]）
语义误解释	模型混淆了语义上相关的晶体学参数因为它缺乏对其不同物理含义的深入理解。	大语言模型错误地将伯格斯矢量识别为晶格常数。（Moon 等人 [44]）
格式 or Unit 换算错误	跳过了必要的单位换算，尽管提示词中有明确的指令。	大语言模型未能换算单位 to 埃 (Å) 在一个 few 个实例。（Piorunek 等人 [45], Jiang 等人 [46]）

至关重要，因为当前评估关注的是在铸态条件下呈现单相体心立方结构的合金。因此，我们将相身份和加工条件整合为复合判据“铸态单相体心立方”，以准确表示这一组合实体。这意味着，如果大语言模型在相或合金加工条件中的任一项上出错，整个条目都会被视为假阳性（FP）或假阴性（FN）。例如，如果大语言模型提取出一种铸态双相（体心立方+ 第二相）合金，却将其误标为单相体心立方，那么整个条目就会被计为假阳性。在计算复合判据“铸态单相体心立方”的精确率和召回率时，该条目会计入假阳性数量；其相关的名义成分和晶格常数也会被归为假阳性，因为它们对应的成分超出了我们的范围或与目标无关。相反，如果大语言模型将铸态单相体心立方合金误分类为双相（体心立方 + 面心立方）合金，那么整个条目就会被计为假阴性。无论其名义成分和晶格常数是否各自被正确提取，都会被归类为目标类别的“未检出”，因为该合金不符合我们的筛选标准。

与来源出版物进行人工核验后发现，所有字段的提取准确性都很高，具体以精确率、召回率和 F1 分数衡量（数据见 表3）。在提取出的 47 条铸态单相体心立方记录中，核验发现了五处错误：三条记录的合金加工条件有误，一条计算结果被误分类为实验铸态数据，还有一条记录被遗漏（FN）。因此，复合标准“铸态单相体心立方”的精确率为 0.915（43/47），召回率为 0.977（43/44）。对于晶格常数字段，我们另外发现了两处错误，因此精确率为 0.872（41/47），召回率为 0.932（41/44）。

名义成分字段的提取性能与复合标准完全一致（精确率 $= 0.915$，召回率 $= 0.977$）。这表明，只要 Claude 3.5 Sonnet 模型正确识别出某条记录属于铸态单相体心立方，就能准确提取其成分。精确率和召回率的计算分母不同：精确率是真正例（TP）占真正例与假正例（FP）之和的比例，而召回率是真正例占真正例与假反例（FN）之和的比例（更多详情见方法部分）。

表3 通过人工核验随机抽取的46篇出版物，计算了 Claude 3.5 Sonnet 大语言模型的精确率和召回率。大语言模型提示词中指定的命名实体包括名义成分、晶格常数、相和合金加工条件。由于我们关注的是经铸态加工且呈现单相体心立方结构的高熵合金成分，因此将相和合金加工条件合并为一个实体：铸态单相体心立方。我们首先筛选提取数据集，仅保留单相体心立方合金，然后在筛选后的数据集上计算精确率和召回率指标。值得注意的是，由于评估实测成分较为复杂，且并非每篇出版物都会报告实测成分，因此评估中未纳入实测成分。

命名实体	精确率	召回率	F1 分数
铸态单相体心立方	0.915	0.977	0.945
名义成分	0.915	0.977	0.945
晶格常数	0.872	0.932	0.901

异常值讨论

在提取出的 17 条具有晶格常数 >10 Å 的记录中，有两篇出版物（Yang 等人 [43] 和 Pan 等人 [47]）并非关于高熵合金，但仍同时出现在通过出版商应用程序接口搜索和 Claude 3.5 Sonnet 相关性分类步骤返回的结果中。Yang 等人 [43]研究了添加少量镍、钛和硅的 G相强化铁素体Fe-Cr模型合金。Claude 3.5 Sonnet 模型似乎因其中含有多种组成元素，便将其误分类为高熵合金，这表明它混淆了多组分合金与高熵合金成分。经进一步检查，我们发现 Claude 3.5 Sonnet 正确记录了合金成分。然而，它误将镍$_{16}$钛$_{6}$硅$_{7}$ 析出相的晶格常数理解为基体的晶格常数。有趣的是，对于这种特定合金，大语言模型能够将单位从纳米换算为 Å，这表明错误并非由格式问题导致，而是因为它在语义上混淆了晶格常数所描述的相。这一讨论是上下文幻觉的一个例子：大语言模型从不同上下文中拼接信息，生成看似合理却不正确的提取结果。我们使用 `EvaluatePdfOutputLoss` 函数确认了这一幻觉。损失函数分析（补充信息中的图 S4）表明，大语言模型正确识别出该晶格常数属于 G相，却错误地将其与

块体合金成分相关联。这表明大语言模型错误地将析出相的晶格常数归于基体合金，未能区分基体相与析出相。

Pan 等人研究中还出现了一个上下文幻觉案例。 [47], 他们研究了一种 CoCrNiVC 中熵合金。Claude 3.5 Sonnet 正确提取了成分、相和加工条件，却错误地将晶格常数报告为 10。9 ˚A，而这个数值对应的是 $Cr_{23}C_6$ 析出相的晶格常数，并非面心立方基体的晶格常数。值得注意的是，文本损失分析（补充信息中的图 S5）显示，Claude 3.5 Sonnet 识别出该晶格常数与碳化物有关。然而，它错误地将其归属于该合金。这表明，尽管模型了解正确的上下文，却未能准确运用这些知识。Wu 等人的研究中也出现了类似错误 [48], ：Claude 3.5 Sonnet 错误地将 11。303 ˚A 提取为高熵合金基体的晶格常数，而这个数值实际描述的是 Ti_2Ni 金属间相。损失函数分析（补充信息中的图 S6）证实了这一错误归属。

在一个案例中，Claude 3.5 Sonnet 模型将柏氏矢量模误认为晶格常数。Moon 等人 [44] 研究了 CoCrFeNiMo 的低温变形行为，但未报告晶格常数。相反，作者报告的柏氏矢量模为 0.255 纳米（nm）。Claude 3.5 Sonnet 模型错误地提取了这一数值（换算为 2.55 ˚A），并将其作为高熵合金晶格常数。这个错误可能源于伯格斯矢量与晶格常数之间的语义相似性。损失函数分析（补充信息中的图 S7）证实模型识别出了这一错误：“不，给出的数值（0.255）似乎是伯格斯矢量模，而不是晶格常数。”

单位换算错误是提取数据集中离群值的一个主要来源。尽管提示指令明确要求以 ˚A 为单位报告晶格常数，但在大规模提取过程中，Claude 3.5 Sonnet 模型经常直接提取原始单位（纳米或皮米）的数值，而不进行换算。以下两个示例说明了这种系统性的格式错误。在 Piorunek 等人 [45], 的论文中，来源报告的晶格常数为 3.19 ˚A，但 Claude 3.5 Sonnet 模型提取的数值为 0.319 ˚A（损失分析见补充信息中的图 S8）。同样，对于 Jiang 等人 [46], 的论文，Claude 3.5 Sonnet 模型提取的数值为 0.323 ˚A，而不是 3.23 ˚A（补充信息中的图 S9）。损失函数的输出揭示了一个令人担忧的模式：Claude 3.5 Sonnet 模型没有承认单位换算错误，反而坚持认为提取的数值（例如，0.319 ˚A、0.323 ˚A）以 ˚A 为单位是正确的。这是一种事后合理化，即模型为错误的提取结果辩解，而非认识到错误。值得注意的是，Jiang 等人的案例还体现了高熵合金相关性分类失败。该论文研究的是 Mg-Gd-Y-Zr 合金，而非高熵合金，却通过了检索和提取筛选器。

测试提示迁移性

为评估提示迁移性在不同模型间的表现，我们在三种近期商用大语言模型上评估了初始未优化提示词和优化后的提示词：(*i*) Claude 4.5 Sonnet、(*ii*) Gemini 2.5 Flash 和 (*iii*) GPT-5。Claude 4.5 Sonnet 和 GPT-5 是针对性能进行优化的模型，而 Gemini 2.5 Flash 则是更廉价的轻量级模型

针对效率和速度进行了优化。在这三个新模型中，Gemini 2.5 Flash 和 GPT-5 默认启用扩展推理；相反，Claude 4.5 Sonnet 的默认设置是禁用扩展推理。Gemini 2.5 Flash 和 GPT-5 中的默认推理参数分别设置为“动态”（思考预算=-1）和“中等”（推理努力程度="medium"）。我们使用三种模型的默认温度设置，在两个不同的出版物集上对其进行了评估：（i）用于提示优化的七篇出版物专家标注训练集（模块 1，图1），其中包含22个手动提取的单相高熵合金条目；以及（ii）随机选取的46篇出版物验证集，用于评估大规模提取性能，其中包含44个经过人工核验的铸态单相体心立方高熵合金条目（上一节已对此进行讨论）。提取性能指标，具体包括精确率、召回率和 F1 分数，分别针对应用于专家标注训练集的初始提示词（补充信息中的表 S1），以及在46篇出版物验证集上评估的优化后的提示词（补充信息中的表S2a和表 S2b）进行估算。三种近期大语言模型（无论是否启用扩展推理）在专家标注训练集上的表现均显著优于 Claude 3.5 Sonnet，即使使用初始未优化提示词也是如此。值得注意的是，使用初始未优化提示词的 Claude 4.5 Sonnet，在专家标注训练集上的表现优于使用完全优化提示词的 Claude 3.5 Sonnet。同样，采用最小推理和初始未优化提示词的 GPT-5，在专家标注训练集上的表现也达到了与使用完全优化提示词的 Claude 3.5 Sonnet 模型相当的水平。

在使用优化提示对包含 46 篇出版物的验证集进行处理时，Claude 4.5 Sonnet 在未启用扩展思考的情况下取得了最佳整体性能，假阳性为零，共有四处错误：漏掉一条记录、一处名义成分错误和两处晶格常数错误。它在所有提取字段上的 F1 分数均为最高，表明在 Claude 3.5 Sonnet 上优化的提示词能够有效泛化至同一大语言模型架构系列中的较新模型。启用扩展推理的 Claude 4.5 Sonnet 表现与默认配置（即未启用扩展推理）相当。召回率基本相同，但由于扩展推理时过度解读间接信息而产生了两个假阳性，精确率有所下降。此外，该模型漏掉了两条记录，并产生一处晶格常数错误。

Gemini 2.5 Flash 和 GPT-5 的表现相近，所有字段的 F1 分数差异均小于 0.04。Gemini 2.5 Flash 提取了 41 条记录，产生两个假阳性；GPT-5 则提取了 42 条记录，产生三个假阳性和一处名义成分错误。尽管架构不同，两种模型的 F1 分数始终较高，与 Claude 系列模型相比，性能仅略有下降。尽管 Gemini 2.5 Flash 是轻量级模型，规模小于 Claude Sonnet 4.5 和 GPT-5，但其表现与 GPT-5 相当，表明提示词能够在不同大语言模型系列之间成功迁移。表 S2a 和表 S2b 中的数据还包括使用各大语言模型处理这组 46 篇出版物的成本（以美元计），以及成本效益比（每美元对应的平均 F1 分数）。比值越高，表示单位成本的提取性能越好。尽管 Gemini 2.5 Flash 的平均 F1 分数相对较低，但其成本效益比最高。

在 Gemini 2.5 Flash 和 GPT-5 产生的假阳性中，一个值得注意的案例涉及 Zhang等人发表的 TiVNbMoTa合金 [49]。在原始出版物中，晶格

常数并未由作者明确报告，而是通过平均原子体积进行间接比较来描述。Claude 3.5 Sonnet 和 Claude 4.5 Sonnet（未启用推理）模型正确识别出该晶格常数并未直接提供，因此没有提取数值。相比之下，Gemini 2.5 Flash（启用动态推理）和 GPT-5（采用中等推理）分别为 TiVNbMoTa合金推断出 3.195 °A 和 3.194 °A 的数值。尽管这些推断值在数学上看似合理，体现出扩展推理能力，但我们仍将其归类为假阳性，因为它们代表的是模型推断，而非直接文本提取。

为进一步确认推理的作用，我们针对这篇特定出版物重新运行了数据提取任务 [49]，使用 Gemini 2.5 Flash 和 GPT-5，并设置了不同的扩展推理预算。对于 Gemini 2.5 Flash模型，在关闭动态推理能力时（即，思考预算=0），模型的回答与出版物一致，并报告晶格常数为“未找到”。¯我们重复了这项测试三次，Gemini 2.5 Flash 模型每次都报告晶格常数为“未找到”。当思考预算设为默认值（动态推理）时，两次独立运行返回了推断结果，另外两次则报告晶格常数为“未找到”。这一结果符合我们的预期：扩展推理可能使回答更难预测，但也更稳健，因为模型会论证多个正确的解答。通过审阅输出中的思维摘要，并将其与推断结果对照，我们进一步了解了推断晶格常数时的推理过程。Gemini 2.5 Flash 模型的扩展推理内容见补充信息中的第 S5 节。Gemini 2.5 Flash 模型首先指出，论文的结果与讨论部分讨论了 TiVNbTa 和 TiVNbMoTa 合金。在后续的推理过程中，可以看到以下句子：“我看到了对两者的比较，其中给出了 TiVNbMoTa 的晶格常数。”

在 GPT-5 模型的最简推理模式下，其表现与采用动态推理的 Gemini 2.5 Flash 相当；四次结果中有两次给出了推断的晶格常数。当 Claude 4.5 Sonnet 模型采用扩展推理且令牌预算为 1024 时，其表现与采用动态推理的 Gemini 以及采用中等推理的 GPT-5 均相近。在四次重新运行中，有一次给出了推断值 3.194 °A，两条思考历史明确详细说明了该数值的计算过程。Claude 4.5 Sonnet 模型扩展推理的一个代表性示例见补充信息的 S6 节。相反，关闭扩展推理后，Claude 4.5 Sonnet 模型针对这篇特定出版物，在三次重新运行中始终返回“未找到”。这证实，推断出的晶格参数是扩展推理过程的结果，而非大语言模型幻觉。

表 S2b 展示了 Claude 4.5 Sonnet 的模型表现，并比较了使用和不使用扩展推理的优化提示词。在扩展推理的令牌预算为 1024 个 token 时，所有实体的 F1 分数略有下降，但总体仍然相当。这一观察表明，使用优化后的提示词时，扩展推理并不会进一步提升数据提取结果。

我们的比较表明，近期扩展思考和推理方面的进展能够显著提升模型在未经优化的提示词下的表现。推理过程旨在分析用户意图，并在无需明确优化提示词的情况下给出预期结果。对于本文研究的相对简单的抽取任务，这种方法证明是有效的；该任务仅涉及四种主要实体（名义成分、合金加工条件、相和晶格常数）。这一结果是否适用于更复杂的数据提取任务（即命名实体的提取），仍需进一步研究，超出本研究的范围。然而，我们的案例研究也凸显了一项关键风险：推理能力更强的模型可能会过度解读间接信息。尽管这些推断值看似合理，但它们反映的是模型的解读，而非对源文本的直接提取。这一发现进一步表明，有必要采用系统化的提示优化框架，例如 >10 TextGrad [26]，以明确设定限制，防止过度解读。这样可以确保先进的推理能力提升而非损害对源材料的忠实度。

总体而言，尽管前沿模型即使使用基础提示词也能达到很高的准确率，但这恰恰进一步说明了系统化提示优化的必要性，而非否定其必要性。Claude 优化提示词将 Gemini 2.5 Flash 的 F1 分数从 0.900 提升至 0.917，且无需重新优化，这表明优化信号能够跨模型系列和模型代际迁移。此外，Gemini 2.5 Flash 在准确率相近的情况下，成本效益是其他模型的四至六倍。这些结果表明，我们的提示优化提炼出了模型无关特征，例如明确无歧义的格式规则、对缺失数据的明确处理方式，以及清晰的范围界定指令。这些特征能够跨模型系列泛化，并可根据可用预算灵活部署。基于这些结果和观察到的失效模式，我们总结了四种进一步提高抽取可靠性的策略。这些策略将在补充信息的 S7 节中讨论。

此外，提示优化会产生一笔固定的一次性成本，在大规模部署中很容易摊销。在我们的工作流程中，优化仅需进行 21 次文章评估（七篇文章，每篇经过三个训练轮次），成本约为完整 2,097 份文档抽取成本的 1%，却使 Claude 3.5 Sonnet 的召回率提高了 3×。对于小规模的探索性任务，直接使用当前最强的模型仍然具有成本效益。然而，对于重复性抽取或大语料库抽取，在能力较强的中端模型上进行优化，再部署到成本较低的模型上，几乎肯定更经济；这样既能达到相近的准确率，又能大幅降低开支。

用于预测晶格常数的机器学习

下一步是开发机器学习模型，以建立铸态单相体心立方合金的晶格常数与高熵合金成分之间的定量关系。补充信息中的图 S13a 将集成支持向量回归（eSVR）模型的预测结果与大语言模型提取的实验测得的晶格常数值进行了比较，显示预测结果与报告的实验数据吻合良好。通过计算构成集成模型的 100 个独立 SVR 预测值的标准差，确定了每种成分的不确定性，

该标准差可作为模型置信度的衡量指标。训练集和 R^2 测试集的R^2分别为 0.989 和 0.969。在晶格常数较低和较高的区域（分别为 ∼2.8–3.0 ˚A 和 ∼3.4-3.6 ˚A）可以观察到一些离散现象。这些范围内较大的不确定性估计表明，与中间范围的数值相比，这些区域的训练样本可能较少，或成分复杂性较高。晶格常数较低区域的合金主要是以 AlCoCrFeNiM$_x$为基础的成分，其中 M 为 Si、 [50], Nb [51],或 Mo [52]。与 eSVR 模型类似，LASSO模型的线性集成（eLASSO）也表现出很强的预测能力。如补充信息中的图S13b所示，预测的晶格常数与实验测得的数值高度吻合。根据 1000 个独立 LASSO估计器的标准差估算出的模型不确定性很小，因此在图中不可见。eLASSO 模型在训练集和测试集上的 $R^2 = 0$R²均达到 0.974，展现出良好的泛化性能。

讨论

我们开发了一种以专家知识为基础、由反馈引导的自动提示优化框架，灵感源自 TextGrad，并以高熵合金晶格常数提取作为代表性测试场景对其进行了验证。仅使用七份人工标注文档，经过三轮训练，该工作流程便将 Claude 3.5 Sonnet 模型的召回率从 0.27 提高至 > 0.90，表明只需极低的成本和人力投入，即可实现有效的反馈引导调优。优化后的提示词可高保真地迁移至不同模型系列（Claude 4.5 Sonnet、GPT-5 和 Gemini 2.5 Flash），在 Gemini 2.5 Flash 上还带来了小幅额外提升，并使其能够部署于成本更低的架构，而无需重新优化。这些结果表明，反馈引导的提示优化是一种可扩展、可迁移且经济高效的方法，可用于未来更复杂的大语言模型辅助科学任务。

将该方法应用于高熵合金晶格常数提取问题后，我们得以从 2,267 篇出版物中大规模检索数据。优化后的提示词对铸态单相体心立方成分的精确率和召回率均超过 0.90，并提取了 1,861 种高熵合金成分的晶格常数数据。基于该语料库，我们整理出一个包含 166 种铸态单相体心立方高熵合金成分的数据集，用于机器学习验证。在这些大语言模型提取的数据上训练的模型，在晶格常数预测方面取得了测试 $R^2 > 0.95$，表明经过适当验证的大语言模型提取数据能够支持定量材料建模，且性能可与人工整理的数据库相媲美。

尽管优化后的提示词能够成功迁移到不同的大语言模型系列中，但我们观察到，推理能力更强的模型有时会通过过度推断引入看似合理但未经报道的数值。这种行为表明，增加推理投入并不能保证认知可靠性，也进一步凸显了系统化验证协议的必要性。因此，提示优化不仅是为了提高召回率，还要约束模型输出，以确保其忠实于源材料。

详细的错误分析进一步揭示了大语言模型的三种典型失效模式：上下文幻觉、语义误解和单位换算错误。这些问题凸显了

在将结果用于定量分析前，仍需进行严格验证。建议采取以下保障措施：成分一致性检查、物理合理性过滤器，以及在下游建模前进行随机抽样验证。展望未来，可通过以下方式进一步提升提取质量：(*i*) 扩大训练文档的多样性，(*ii*) 探索有限形式的输出层面优化，(*iii*) 优先使用出版商原始PDF，而非XML重构文件；以及(*iv*) 定期使用当前一代最强大的前沿模型重新评估性能。有关如何基于本研究提取的数据开展任何下游任务的更多指令，请参见补充信息中的S9节。

总的来说，本研究表明，提示优化只需付出极少的成本和努力，便能显著提升性能；优化后的提示词可跨架构泛化，但需要经过验证，以防止推理导致过度推断；经过验证的 LLM 提取数据集则有助于在材料科学领域开展高保真度的科学数据挖掘。

方法

专家指导的提示优化工作流

专家指导的基于LLM的数据提取工作流如图1所示。该方法由三个模块组成。第一个模块包括对选定文献进行专家标注，以定义目标数据结构和提取模式，为提示工程提供依据。例如，为提取晶格常数，我们为每种高熵合金标注以下信息：合金名称、名义成分、实测成分、相、合金加工条件（如铸态、退火态或粉末加工）以及相应的晶格常数。这些信息汇编自少量研究论文。在本研究中，我们随机选取了七篇出版物，并提取了22个数据点。

第二个模块通过TextGrad 框架 [26]执行自动化的迭代式提示优化。提示优化流程的伪代码见补充信息中的图S14。我们从描述抽取任务的初始自然语言提示开始，调用兼容PDF的多模态LLM（Claude- 3.5-Sonnet-20240620） [35] ，从模块1中由领域专家标注的相同出版物中抽取结构化数据。抽取质量通过自定义损失函数 `EvaluatePdfExtractionLoss` 进行评估，该函数计算大语言模型生成的输出与模块1中的专家整理的真实数据集之间的差异。`EvaluatePdfExtractionLoss`函数接受四个输入：当前提示词、PDF 格式的科学出版物、专家构建的数据，以及大语言模型生成的对应数据。损失函数使用定制化评估提示（详见补充信息中的图S15），指示大语言模型评估生成的输出是否同时符合抽取提示的要求和PDF内容。完成分析后，该函数会明确说明生成的数据是否与提示词一致。该函数生成的反馈随后由反向引擎用于修改提示词，从而实现迭代优化过程。在本研究中，根据该模块的预算，优化进行了三个

训练轮次。为确保不超过令牌限制，出版物按每组三篇进行批处理。

第三个模块使用模块 2 中的优化后的提示词，对高熵合金文献开展大规模数据提取。我们应用该提示词提取了以下命名实体：合金名称、名义成分、实测成分、相、合金加工条件（例如铸态、退火态、粉末加工）以及晶格常数。我们共使用了从多个来源获取的 2,267 篇出版物，其中包括通过爱思唯尔应用程序编程接口（应用程序编程接口）获取的全文 XML 文件，以及其他出版商提供的 PDF 文档。更具体地说，我们根据 Polak 和 Morgan 的工作中报告的数字对象标识符，获取了 2,093 个全文 XML 文件并将其重构为 PDF（文本 + 表格，无图），[10], 此外，还从各出版商处下载了 174 份完整 PDF（文本 + 表格， + 图）。在模块 2 中，Claude 3.5 Sonnet 模型按照 TextGrad 中的规定，对正向引擎和反向引擎均使用默认温度值 0.0。在模块 3 中，我们使用 Claude 3.5 Sonnet 模型的默认温度值 1.0。同样，对于 Claude 4.5 Sonnet 、 Gemini 2.5 Flash 和 GPT-5 模型，我们也保持默认温度值 1.0。

数据提取性能评估

我们的工作流程需要采用两种独立的评估方案来评估提示词的表现：（i）使用提示优化期间采用的专家依据参考数据集进行验证，以及（ii）在未见过的出版物上部署后，评估大规模提取的质量。为定量评估性能，我们基于补充信息中表 S3 给出的混淆矩阵，使用 公式 1、公式 2 和 公式 3 分别计算精确率、召回率和 F1 分数 [53, 54]。

$$\text{Precision} = \frac{\text{Relevant Retrieved}}{\text{All Retrieved}} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{1}$$

$$\text{Recall} = \frac{\text{Relevant Retrieved}}{\text{All Relevant}} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{2}$$

$$\text{F1 Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

我们的高熵合金数据提取工作流程分为两个层级：（i）材料识别和（ii）属性提取。只有两个层级都正确，检索结果才会被归类为真正例（TP）：材料必须是论文中讨论的高熵合金，并且其性能必须得到准确提取。如果识别出的材料与论文中提及的高熵合金不符，则所有相关参数都归类为假正例（FP）。即使提取的性能值准确无误，这一分类也依然成立。

除了精确率和召回率，我们还进行了一致性检查，以在下游任务开始前验证提取质量。这些合理性检查用于核实三个相关字段之间的一致性：合金名称、名义成分和实测成分。这些字段在组成元素方面应保持一致，且

原子比应大致一致，因为它们是对同一材料的不同表示方式。我们使用 L1距离（D_{L1} 见公式 4）和余弦相似度（CosSim，见公式 5）量化成分字段之间的一致性。在这些公式中，$\mathbf{v}_1$ 和$\mathbf{v}_2$ 分别表示待比较两个字段（例如名义成分与实测成分）的成分向量，其中每个元素 i 对应于任一成分中存在的 n 种不同组成元素之一。

$$D_{\mathrm{L1}}(\mathbf{v}_1, \mathbf{v}_2) = \sum_{i=1}^{n} |v_{1,i} - v_{2,i}|\,, \tag{4}$$

$$\mathrm{CosSim}(\mathbf{v}_1, \mathbf{v}_2) = \frac{\sum_{i=1}^{n} v_{1,i} v_{2,i}}{\sqrt{\sum_{i=1}^{n} v_{1,i}^2}\sqrt{\sum_{i=1}^{n} v_{2,i}^2}}. \tag{5}$$

计算 D_{L1} 和 CosSim 后，我们手动检查了合金名称、名义成分和实测成分字段中成分差异较大的条目（ D_{L1} 较大、CosSim 较低），因为这些不一致可能表明材料识别错误或数据关联有误。此外，我们剔除了晶格常数大于 10 °A 或小于 1 °A 的条目，这些问题可能是大语言模型进行数据提取时发生单位换算错误所致。

我们定义了第二个文本损失函数 `EvaluatePdfOutputLoss`，用于评估提取结果的忠实度。受 Polak 和 Morgan 的观察启发 [10]，即后续提问可以提高提取质量，我们在损失函数中将若干问题设为验证指令（见补充信息中的图 S16）。该函数以大语言模型生成的输出和原始 PDF 作为输入，然后通过事实核查查询评估准确性，例如："输出中的晶格常数确实是该材料的晶格常数，而不是其他材料的吗？"。我们主要使用此损失函数诊断异常值数据中的错误，详见补充信息中的 S3 节。

下游任务：用于预测晶格常数的机器学习

经过数据清理和质量控制后，筛选出159种单相体心立方高熵合金成分，用于后续机器学习任务。对于每种成分，我们使用 Magpie tool [55], 计算描述符，该工具生成元素性质（例如原子体积、共价半径）的成分加权平均值。我们采用基于递归特征消除方法 [56–58] 的描述符筛选流程，识别用于训练机器学习模型的最重要描述符（见补充信息中的 S13 节）。特征筛选确定了六个成分加权描述符，包括：原子体积（"meanAtomicVolume"）、共价半径（"meanColaventRadius"）、门捷列夫数（"meanMendeleev"）、电负性（"meanElectronegativity"）、 d轨道中的价电子数（"meanNdValence"），以及未填充价轨道数（"meanNUnfilled"）。

将包含159种成分的数据集按80:20的比例划分为训练集和测试集（127个训练样本，32个测试样本）。我们训练了两种机器学习模型：（i）由采用径向基函数（RBF）核函数 [59–61]的非线性支持向量回归模型组成的集成支持向量回归（eSVR）模型，以及（ii）线性最小绝对收缩与选择算子（eLASSO）模型集成 [62, 63]。通过网格搜索，分别优化集成中各支持向量回归估计器的超参数（gamma 和 cost）。根据测试集上的最高 $\{10, 20, 30, 40, 50, 75, 100\}$ 性能，从若干估计器中 R^2 选择最优集成规模。eLASSO 模型使用从训练数据生成的 1,000 个自助法重采样样本进行训练。对于每个重采样样本，通过网格搜索优化正则化参数（收缩惩罚项），选择使 10 折交叉验证误差最小的取值。eSVR 和 eLASSO 使用相同的训练集与测试集划分，以便直接比较性能。

补充信息。 [URL] 将由出版商插入。

资助。 S.L. 和 P.V.B. 感谢弗吉尼亚大学研究创新奖提供的资助。

致谢。 S.L. 和 P.V.B. 感谢弗吉尼亚大学研究创新奖提供的资助。Y.J. 和 P.V.B. 感谢弗吉尼亚大学研究兴趣小组倡议促进合作与思想交流。S.L.、T.R.B. 和 P.V.B 感谢 Sunidhi Garg 和 Yifei Duan 对稿件提出的宝贵意见。

作者贡献

本研究由 S.L.、Y.J.、W.R. 和 P.V.B. 设计。稿件由 S.L.、T.R.B.、Y.J.、W.R. 和 P.V.B. 撰写。专家构建的数据集由 S.L. 和 T.R.B. 在 P.V.B. 的指导下创建。S.L.、T.R.B. 和 P.V.B. 收集了用于数据提取的文献。数据提取工作流程由 S.L. 在 Y.J.、W.R. 和 P.V.B. 的指导下开发。S.L.、T.R.B. 和 P.V.B. 对提取的数据进行了人工核验。所有作者均参与了结果讨论、论文撰写，并对稿件提供了反馈。

利益冲突

作者声明不存在利益冲突。

数据可用性

专家构建的数据集、提取的晶格常数数据集、用于评估的抽样数据集（见表3）、提示词和文本损失均已存档于GitHub：https://github.com/Shunl1996/expert-paper-extract，并随论文一并提供。用于大规模数据提取的论文数字对象标识符以数据集形式提供。每篇论文对应的大语言模型原始JSON输出，以及部分论文的思维摘要，均已存档于 UVA Dataverse [64]，以确保可复现性。

代码可用性

提示优化、大规模数据提取和成分验证的代码可在以下网址获取：
https://github.com/Shunl1996/expert-paper-extract，也与上述数据一并存档。

参考文献

[1] Kim, E., Huang, K., Saunders, A., McCallum, A., Ceder, G., Olivetti, E.：通过文本提取和机器学习从科学文献中获取材料合成洞见。《材料化学》29(21)，9436–9444（2017）[2] Venugopal, V., Broderick, S.R., Rajan, K.：一图胜千言：应用自然语言处理工具创建量子材料数据库图谱。《MRS通讯》9(4)，1134–1141（2019）[3]Olivetti, E.A., Cole, J.M., Kim, E., Kononova, O., Ceder, G., Han, T.Y.-J., Hiszpanski, A.M.：自然语言处理和信息提取赋能的数据驱动的材料研究。《应用物理评论》7(4)（2020）[4]Shetty, P., Ramprasad, R.：使用自然语言处理从聚合物文献中自动提取知识。Iscience 24(1)（2021）[5] Wang, W., Jiang, X., Tian, S., Liu, P., Dang, D., Su, Y., Lookman, T., Xie, J.：通过文本挖掘构建高温合金数据的自动化流程。npj Computational Materials 8(1)，9（2022）[6] Zheng, Z., Zhang, O., Borgs, C., Chayes, J.T., Yaghi, O.M.：用于文本挖掘和金属有机框架合成预测的 ChatGPT 化学助手。《美国化学会志》145(32)，18048–18062（2023）[7] Gilligan, L.P., Cobelli, M., Taufour, V., Sanvito, S.：从科学文献中自动生成数据库的无规则工作流。npj Computational Materials 9(1)，222（2023）[8] Sayeed, H.M., Mohanty, T., Sparks, T.D.：标注材料科学文本：一种利用 Gemini Pro 制作输出的半自动化方法。Integrating Materials and Manufacturing Innovation 13(2)，445–452（2024）[9] Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A.S., Ceder, G., Persson, K.A., Jain, A.：利用大语言模型从科学文本中进行结构化信息提取。《自然通讯》15(1)，1418（2024）[10] Polak,M.P., Morgan,D.：利用对话式语言模型和提示工程从研究论文中提取准确的材料数据。《自然通讯》15(1)，1569（2024）https://doi.org/10.1038/s41467-024-45914-8

[11] Jiang, X., Wang, W., Tian, S., Wang, H., Lookman, T., Su, Y.：自然语言处理和大语言模型在材料发现中的应用。npj Computational Materials **11**（1），79（2025）[12] Foppiano, L., Lambard, G., Amagasa, T., Ishii, M.：利用大语言模型从材料科学文献中挖掘实验数据：一项评估研究。Science and Technology of Advanced Materials: Methods **4**（1），2356506（2024）[13] Liu, S., Wen, T., Pattamatta, A.S., Srolovitz, D.J.：提示工程化大语言模型与深度学习相结合的材料分类工作流。《材料》 **80**，240–249（2024）[14] Ekuma, C.E.：利用对话模型进行动态上下文学习，以实现数据提取和材料性质预测。APL Machine Learning **3**（1），016119（2025）[15] Biswajeet, D.D., Kadkhodaei, S.：利用大语言模型解决机器学习中的数据稀缺问题：石墨烯合成的应用。arXiv 预印本 arXiv:2503.04870（2025）[16]Liu, N., Jafarzadeh, S., Lattimer, B.Y., Ni, S., Lua, J., Yu, Y.：利用大语言模型实现聚合物性质的数据稀缺学习。Nature Computational Science **5**（3），245–254（2025）[17] Turan, I., Sparks, T.：革新科学图表解读：大语言模型数据提取性能的基准测试。ChemRxiv。2025（2025）https://doi.org/10.26434/chemrxiv-2025-pcq8d[18] Polak, M.P., Morgan, D.：利用多模态大语言模型的视觉能力从绘图中自动提取数据。arXiv 预印本 arXiv:2503.12326（2025）[19] White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., Schmidt, D.C.：提示模式目录：通过 ChatGPT 提升提示工程。arXiv 预印本 arXiv:2302.11382（2023）[20] Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., 等：提示报告：提示工程技术的系统综述。arXiv 预印本 arXiv:2406.06608（2024）[21]Giray, L.：使用 ChatGPT 进行提示工程：学术写作者指南。Annals of Biomedical Engineering **51**（12），2629–2633（2023）[22] Ha, E., Ha, S.M., Gerelkhuu, Z., Kim, H.-Y., Yoon, T.H.：基于人工智能的纳米毒性数据提取与纳米毒性预测。Computational and Structural Biotechnology Journal **29**，138–148（2025）[23] Zhang, Y., Itani, S., Khanal, K., Okyere, E., Smith, G., Takahashi, K., Zang, J.：

GPTArticleExtractor：用于构建磁性材料数据库的自动化工作流程。《磁性与磁性材料杂志》597，172001（2024）[24] Gupta, S., Mahmood, A., Shetty, P., Adeboye, A., Ramprasad, R.：使用大语言模型从聚合物文献中提取数据。Communications Materials 5(1)，269（2024）[25] Prasad, D., Pimpude, M., Alankar, A.：利用大语言模型开发材料工程自动化知识图谱和数据库。arXiv 预印本 arXiv:2402.11323（2024）[26] Yuksekgonul, M., Bianchi, F., Boen, J., Liu, S., Lu, P., Huang, Z., Guestrin, C., Zou, J.：通过反向传播语言模型反馈优化生成式人工智能。《自然》639(8055)，609–616（2025）[27] Miracle, D.B., Senkov, O.N.：高熵合金及相关概念的批判性综述。《材料学报》122，448–511（2017）[28] George, E.P., Curtin, W.A., Tasan, C.C.：高熵合金：力学性能与变形机制专题综述。《材料学报》188，435–474（2020）[29] Han, L., Zhu, S., Rao, Z., Scheu, C., Ponge, D., Ludwig, A., Zhang, H., Gutfleisch, O., Hahn, H., Li, Z., 等：多功能高熵材料。《自然综述：材料》9(12)，846–865（2024）[30] Borg, C.K., Frey, C., Moh, J., Pollock, T.M., Gorsse, S., Miracle, D.B., Senkov, O.N., Meredig, B., Saal, J.E.：多主元合金力学性能及观测相的扩展数据集。《科学数据》7(1)，430（2020）[31] Tandoc, C., Qi, L., Hu, Y.-J.：用于准确预测体心立方固溶体合金晶格参数的键基模型。Materialia 40，102410（2025）[32] Maresca, F., Curtin, W.A.：难熔体心立方高熵合金在最高1900 K时高强度的机理起源。《材料学报》182，235–249（2020）[33] Liu, S., Lee, K., Balachandran, P.V.：结合机器学习与机理模型预测高熵合金的屈服强度。Journal of Applied Physics 132(10)（2022）[34] Li, J., Fu, X., Feng, H.,Liu, B., Liaw, P.K., Fang, Q.：评估多主元合金中固溶体、局部化学有序和析出强化的贡献。《合金与化合物杂志》938，168521（2023）[35]Anthropic：Claude 3.5 Sonnet。https://www.anthropic.com/claude。大语言模型（2024）[36] Feng, R., Feng, B., Gao, M.C., Zhang, C., Neuefeind, J.C., Poplawsky, J.D.,

Ren, Y., An, K., Widom, M., Liaw, P.K.：过饱和难熔高熵合金具有优异的高温强度。Advanced Materials 33(48)，2102401（2021）[37]Lee, K., Balachandran, P.V.：用于预测铸态多主元合金屈服强度的可解释人工智能方法。Materialia 26，101628（2022）[38] Karumuri, S., Hernandez, A., Mishra, S., McClure, Z., Tucker, V., Flanagan, J.C., Hwang, S., Sandhage, K.H., Bilionis, I., Titus, M.S., 等：结合物理学、机器学习和实验设计高硬度复杂浓缩合金。Journal of Applied Physics 138(8)（2025）[39] Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., 等：人工智能海洋中的塞壬之歌：大语言模型幻觉综述。Computational Linguistics，1–46（2025）[40] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., 等：大语言模型幻觉综述：原理、分类体系、挑战与开放问题。ACM Transactions on Information Systems 43(2)，1–55（2025）[41] Lin, C.-M., Juan, C.-C., Chang, C.-H., Tsai, C.-W., Yeh, J.-W.：添加铝对难熔 $Al_xHfNbTaTiZr$ 合金力学性能和微观结构的影响。《合金与化合物杂志》624，100–107（2015）https://doi.org/ 10.1016/j.jallcom.2014.11.064[42] Zuo, T.T., Li, R.B., Ren, X.J., Zhang, Y.：添加铝和硅对 CoFeNi等原子比合金结构与性能的影响。《磁性与磁性材料杂志》371，60–68（2014）https://doi.org/10.1016/j.jmmm.2014. 07.023[43] Yang, M., Huang, C., Han, J., Wu, H., Zhao, Y., Yang, T., Jin, S., Wang, C., Li, Z., Shu, R., Wang, C., Lu, H., Sha, G., Liu, X.：通过调控晶内和晶界G相析出，开发高强度、高延展性铁素体合金。Journal of Materials Science & Technology 136，180–199（2023）https://doi.org/10.1016/j.jmst.2022.07.029[44] Moon, J., Tabachnikova, E., Shumilin, S., Hryhorova, T., Estrin, Y., Brechtl, J., Liaw, P.K., Wang, W., Dahmen, K.A., Zargaran, A., Bae, J.W., Do, H.-S., Lee, B.-J., Kim, H.S.：Co-Cr-Fe-Ni-Mo中熵合金在极低温下的变形行为。Materials Today 50，55–68（2021）https: //doi.org/10.1016/j.mattod.2021.08.001[45] Piorunek, D., Frenzel, J., J¨ons, N., Somsen, C., Eggeler, G.：高熵形状记忆合金中的化学复杂性、微观结构与马氏体相变。Intermetallics 122，106792（2020）https://doi.org/10.1016/j.intermet.

2020.106792

[46] Jiang, Y., Tang, H., Li, Z., Zheng, D., Tu, Y.：增材制造Mg-Gd-Y-Zr合金：Gd含量对微观结构演变和力学性能的影响。《增材制造》**59**，103136（2022）https://doi.org/10.1016/j.addma.2022.103136[47]Pan, S., Zhang, J.Q., He, B.B., Huang, M.X.：层片状碳化物对CoCrNiVC中熵合金中高应变梯度诱导纳米晶粒的影响。《材料科学与工程：A辑》**852**，143692（2022）https://doi.org/10.1016/j.msea.2022.143692[48] Wu, C.L., Zhang, S., Zhang, C.H., Zhang, H., Dong, S.Y.：通过激光表面合金化在304不锈钢上制备的fecocralniti $_{t}extrmx$ 高熵合金涂层的相演变及空化侵蚀-腐蚀行为。《合金与化合物杂志》**698**，761–770（2017）https://doi.org/10.1016/j.jallcom.2016.12.196[49] Zhang, F.X., Tong, Y., Kirkham, M., Huq, A., Bei, H., Weber, W.J., Zhang, Y.：难熔体心立方固溶合金的结构无序、相稳定性和压缩性。《合金与化合物杂志》**847**，155970（2020）https://doi.org/10.1016/j.jallcom.2020.155970[50] Zhu, J.M., Fu, H.M., Zhang, H.F., Wang, A.M., Li, H., Hu, Z.Q.：多主元AlCoCrFeNiSix合金的合成与性能。《材料科学与工程：A辑》**527(27)**，7210–7214（2010）https://doi.org/10.1016/j.msea.2010.07.049[51] Ma, S.G., Zhang, Y.：Nb添加对AlCoCrFeNi高熵合金微观结构和性能的影响。《材料科学与工程：A辑》**532**，480–486（2012）https://doi.org/10.1016/j.msea.2011.10.110[52] Zhu, J.M., Fu, H.M., Zhang, H.F., Wang, A.M., Li, H., Hu, Z.Q.：多组元AlCoCrFeNiMo$_x$ 合金的微观结构和压缩性能。《材料科学与工程：A辑》**527(26)**，6975–6979（2010）https://doi.org/10.1016/j.msea.2010.07.028[53] Wang, X., Raj, A., Luebbe, M., Wen, H., Xu, S., Lu, K.：利用带有来源追踪的多阶段大语言模型，从文献中可靠地端到端提取材料信息（2025）。https://arxiv.org/abs/2510.05142[54] Manning, C.D.,Raghavan, P., Sch¨utze, H.：《信息检索导论》。剑桥大学出版社，纽约（2008）[55]Ward, L., Agrawal, A., Choudhary, A., Wolverton, C.：用于预测无机材料性能的通用机器学习框架。npj Computational Materials **2(1)**，16028（2016）

[56] Wang, J., Xu, P., Ji, X., Li, M., Lu, W.：钙钛矿材料设计与发现中的机器学习特征选择。《材料》16（2023）[57] Sheikh, S., Vela, B., Attari, V., Huang, X., Morcos, P., Hanagan, J., Acemi, C., Karaman, I., Elwany, A., Arroyav´e, R.：探索化学与增材制造设计空间：关于可打印合金计算引导设计的观点。《材料研究快报》12，235–263（2024）[58] Farrar, E.H.E., Grayson, M.N.：机器学习与半经验计算：协同实现快速、准确且基于机理的反应能垒预测。《化学科学》13，7594–7603（2022）[59] Smola, A.J., Sch¨olkopf, B.：支持向量回归教程。《统计与计算》14(3)，199–222（2004）[60] Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F.：e1071：统计系概率论组（原E1071）的杂项函数，维也纳工业大学。R软件包版本1.6-7。http://CRAN.R-project.org/ package=e1071[61] Balachandran, P.V.：面向高效材料设计的自适应机器学习。《MRS通报》45(7)，579–586（2020）[62] Tibshirani, R.：通过套索回归实现回归收缩与选择。《英国皇家统计学会志B辑：统计方法学》58(1)，267–288（1996）[63] Holleis, L.,Shivaram, B.S., Balachandran, P.V.：机器学习指导的单分子磁体设计及其磁热应用。《应用物理快报》114(22)，222404（2019）https://doi.org/10.1063/1.5094553[64] Liu, S., Booth, T.R., Ji, Y., Reinhart, W., Balachandran, P.V.：复现数据：利用大语言模型从科学出版物中提取高熵合金晶格常数的专家指导的提示工程。https://doi.org/10.18130/V3/ABITMJ . https://doi.org/10.18130/V3/ABITMJ