arXiv:2611.29445v1 [cond-mat.supr-con] 22 Nov 2026

审计由大语言模型构建的超导数据库

Wenjian Lu[1], *[1]*Key Laboratory of Materials Physics, Institute of Solid State Physics, HFIPS, Chinese Academy of Sciences, Hefei 230031, China*（日期：2026年11月24日）

超导材料的数据以测量值的形式发表，并散见于文献中的正文、表格和图表里，因此要将这些数据汇集到一处，就意味着必须阅读论文原文。大语言模型（LLM）能够阅读这些论文，以这种方式构建的数据库也开始出现。难点不在于阅读，而在于可信度：从句子中提取出的数值可能属于相邻的化合物，计算得到的转变温度可能在未注明其为计算值的情况下被记录，而论文原文自身也可能存在错误。本文介绍了一种处理流程，利用大语言模型从已发表文献中抽取超导材料，将每个数值与其来源进行核对，并探究这种核查能够达到何种程度。该处理流程提出了19,576条记录；一系列门控条件拒绝了659,252项字段级抽取，其中644,438项是因为缺少逐字证据，13,121项是因为文本没有将数值归属于该材料，另有1,693项是因为来源句子的表述显示其为计算结果。最终保留下来的数据库包含3,137条材料记录，每条记录都附有其数值所来源的句子。由两个独立模型对幸存项进行审计，并按数据库中的比例抽取500条记录进行分析，结果表明，这些门控条件能够排除机械性错误：在经审计的记录中，93.4%的记录其数值确实出现在文本中，且文本讨论的正是该记录所标注的化合物；其中91.8%的数值属于测量值而非预测值。

关键词：超导性 、大语言模型、材料数据库、数据提 取、数据库整理

I. 引言

超导性是一种材料的电阻在低于转变温度 T_c时消失的现象，而相关测量值自首次观测 [1] 以来便陆续发表于论文中。其结果是，超导体的数据散见于各类期刊的正文、表格和图表中，汇集这些数据的资料规模有限。NIMS SuperCon数据库通过人工收集转变温度 [2]；晶体学开放数据库收录晶体结构 [3]；Materials Project则提供数百万种化合物的计算性质 [4]。这些数据库各有价值，但其涵盖范围都受限于投入其中的工作。

文献本身并不受此类限制。从已发表文本中提取材料及其性质由来已久 [5], ，而大语言模型的出现已将其转变为自动化任务：模型阅读论文并返回结构化记录，如今人们正以这种方式构建磁性 [6, 7] 和热电 [8] 材料数据库。超导性文献也已采用同样的方法进行挖掘 [9, 10], ，近期还整理出了一个包含晶体结构的超导体数据集 [11]。

这种方法在各个领域都存在同样的弱点，而这并非模型本身的弱点：问题在于轻信模型。模型读取一段文字后，可能会报告相邻化合物的转变温度，或给出第一性原理论文中的数值，却不说明该数值是计算得出的；也可能给出论文从未提及的数值。显而易见的补救办法是逐一核对数值及其来源，而接下来的显而易见的问题是，这样做效果如何。答案无法从处理流程中看出：处理流程只能看到它保留下来的内容。大语言模型构建的数据库若经过核查，核查者也只是另一个大语言模型，它读取的是同样的句子 [7] ——这样做确有帮助，但并不等同于由了解该领域的人来核查。

本文介绍了一种处理流程：利用大语言模型从已发表文献中抽取超导材料，并衡量其保留内容的可信度。三项设计使这一衡量成为可能。第一，对每个抽取出的数值而非每篇论文应用一系列级联核查，包括逐字证据、材料绑定、物理合理性、数值的获取方式以及跨论文一致性。第二，数据库会保留每个数值所依据的原句，因此读者、另一个模型或专家都能对照来源核查任何记录，无需重新运行抽取流程。第三，审计由两位评审者完成，他们都不是执行抽取的模型，并且每位评审者都会针对每个数值回答相同的三个问题。

II. 处理流程

A. 从 PDF 到记录

从出版商处收集论文，使用 docling 将其转换为文本 [12], ，再通过一个提示词处理成分，并通过七个特定维度提示词（而不是一个提示词）让大语言模型（MiniMax-M3.1-Flash）读取：基础超导性质、压力依赖关系、同位素与 Uemura 分析、比热、费米参数和电阻率、晶体结构，以及

* 通讯作者： wjlu@issp.ac.cn

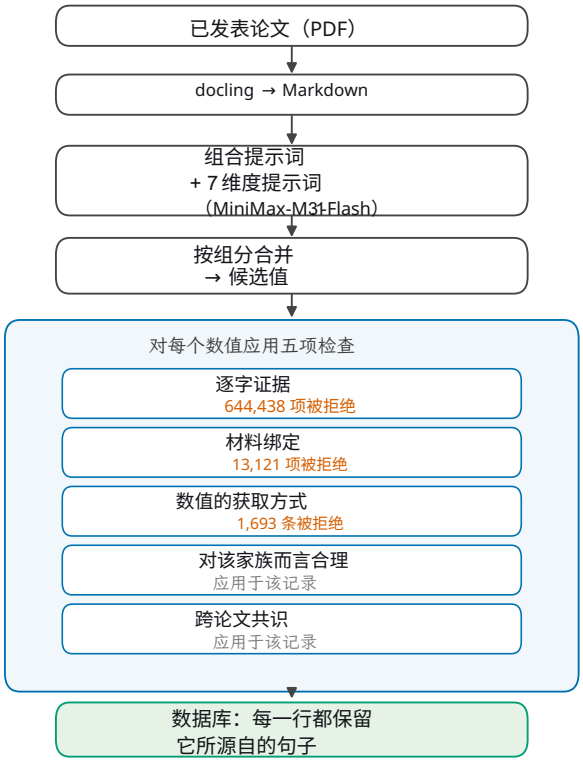

图1. 处理流程与五道关卡。各项计数表示每道关卡拒绝的字段级抽取数量，占语料库抽取总量的比例；两道不会在字段级拒绝任何内容的关卡，则是针对整条记录进行检查。

制备。维度拆分对下文的讨论至关重要，因为它使证据核查能够按字段进行：一篇报告了 T_c 但没有报告 H_{c2} 的论文，可以从中提取 T_c ，但不能提取 H_{c2} 。

维度结果合并为每种组分一条记录；论文提供相关数据时，记录会包含组分、超导性质以及晶体和制备数据。目前尚未将任何内容写入数据库。

B. 五道关卡

合并后保留下来的数值即为候选值，每个候选值都要经过五项检查。

逐字证据。 该数值必须出现在论文中，并且拼写方式必须与论文中的写法一致。这项检查失败的频率高于其他任何检查，而其失败原因与模型无关：大多数性质并不出现在大多数论文中。比较时会合并排版变体，因此“$T_c = 39$K”和“39 K”是同一个数值，但“39 K”仍然不等于“39.5 K”。

材料绑定。 承载数值的句子也必须提及材料本身，或其归一化形式。超导体论文常在同一段中列出许多化合物，模型在阅读时可能会把相邻化合物的

表 I。在对语料库进行抽取时，被门控条件拒绝的字段级抽取。统计的是单个值，而不是记录：一条记录最多提供十个性质。每篇论文只统计一次，取其结果进入数据库的那次运行；被抽取多次的论文，按其最后一次运行统计。

字段	拒绝	原因
T_{CDW}	79,323	无逐字证据
H_{c1}	79,341	没有逐字证据
J_c	78,490	没有逐字证据
穿透深度	78,233	没有逐字证据
相干长度	77,207	无逐字证据
H_{c2}	76,419	无逐字证据
能隙	75,955	无逐字证据
T_c	53,579	无逐字证据
配对对称性	23,213	无逐字证据
机制	22,678	没有逐字证据
T_c	13,121	未说明该材料的相关信息
T_c	1,693	一次计算
总计	659,252	

T_c 关联到句子的主语。这项检查可以防止将“FeSe”记录为 109 K——这是同一段中报告的、基底上单层材料的转变温度，而非块体转变温度。

物理合理性。 每个家族都有记录在案的可接受转变温度范围，超出范围的数值会被拒绝。这项检查刻意采取保守做法：拒绝一项真实测量只会损失一条记录，而接受一项错误归属的测量则会悄无声息地污染数据库。

数值的获取方式。 论文常在同一段中同时引用预测的和观测到的转变温度，而醒目的理论数值往往占据上风。分类器会重新阅读数值所在句子的上下文：预测性措辞会使该数值被排除，观测性措辞则会使其保留。若措辞足以定论，记录便依此处理；若句中没有任何相关措辞，记录就沿用抽取模型对该数值性质的判断，而数据库中的标记记录的正是这一判断。

跨论文一致性。 多篇论文报告的同一种化合物会对应多个数值，而单靠任何一项检查都无法发现某项测量虽出自论文自身、却不适用于该材料的情况。彼此差异在容差范围内的数值会归并到同一个簇中；若没有任何簇占据主导地位，则保留记录更详尽的那条记录。

C. 门控条件拒绝的内容

表I 统计了每个筛选门剔除的内容。这个模式正是重点：罕见性质被拒绝的频率远高于常见性质

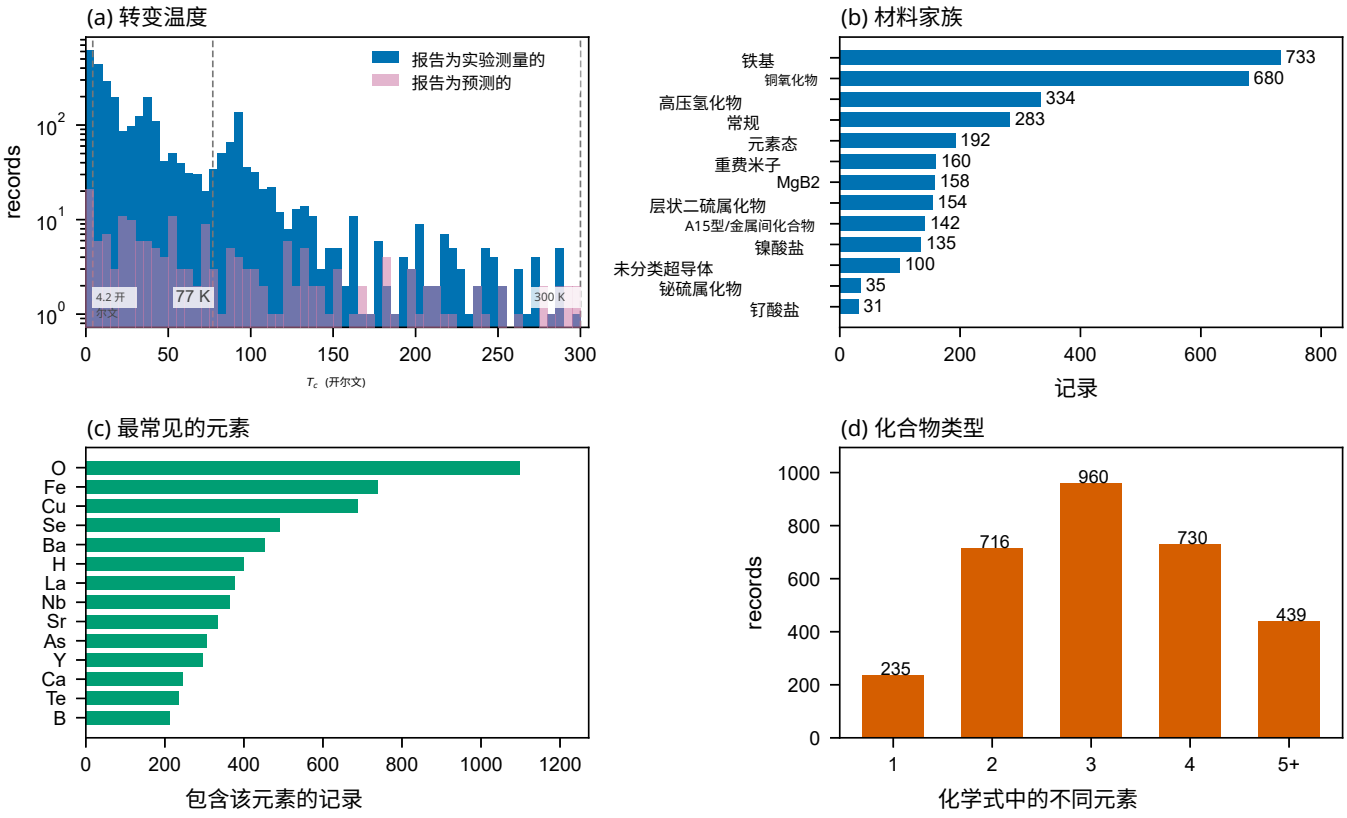

图 2. 提取的超导材料语料库。(a) 转变温度的分布；来源报告的转变温度为预测值的记录以半透明方式叠加在报告为观测值的记录之上；虚线标示液氦（4.2 K）、液氮（77 K）和室温（300 K）的温度。(b) 材料家族。(c) 最常见的元素。(d) 化学式中不同元素的数量。数据共包含 3,137 条记录，涉及 1,371 种不同组分，来源为 26,210 篇论文（2026年11月21日）。

。 T_{CDW}、 H_{c1}、 J_c、穿透深度、相干长度、 H_{c2} 和能隙各自损失了超过 75,000 条字段级抽取，而T_c——几乎每篇超导体论文都会提及的性质——损失了 53,579 条。配对对称性和机制通常以文字而非数字报告，损失的条目更少。

最后两行中的三种原因值得区分，因为它们代表不同的失败。没有逐字证据的数值，很可能从未出现在论文中。文本未将某个数值与材料联系起来，说明该数值是从错误的位置读出的：句子提到的是另一种化合物。文本称某个数值是通过计算得出的，这根本不属于阅读错误：论文正确地陈述了该数值，也说明了它的性质。若承载该数值的句子表明它是计算结果，这类数值就会被剔除；这正是这 1,693 个数值的情况。得以保留的计算值，其所在句子则没有明确表述数值的性质；在这种情况下，记录会保留抽取模型对该数值性质的判断，使两类数值之间的区别得以留存于数据库中。图1 展示了这一筛选流程，以及最终进入表格的数量。

III. 数据库

该语料库是由 26,210 篇论文构成的一个快照。我们已对其中 26,184 篇进行了抽取，并提出了 19,576 条记录；其中 8,076 条所列的组分并非超导体，这是模型读取参考文献列表时典型的幻觉。在通过这一筛选的 11,500 条记录中，7,090 条成为新行，其余则合并到已有行中，或因其证据句表明相应转变温度是计算得到的而被剔除。跨论文整合并去重后，最终得到 3,137 条记录，涵盖 1,371 种不同的组成，且每条记录都附有其数值所依据的句子。

图2 展示了语料库的内容。转变温度范围为 0.01 K 至 298 K，跨越四个半数量级；其中有五条记录的温度为 0 K。其分布与文献所呈现的一致：30 K 以下密集分布，高T_c 温化合物则形成一条长尾。材料族也都是熟悉的那些（以铁基和铜氧化物材料为主，还有相当规模的氢化物族），另有 100 条记录仍未分类；这是抽取方面的事实，而非物理学方面的事实。略多于一半

表 II。对保留记录进行的两次审计。两位评审者都被要求回答相同的三个问题，因此他们的比率具有可比性，且文本逐条记录进行比较。比率附有 95% 威尔逊区间。第三个问题询问数值是如何获得的，而不是该记录是否成立：论文为正确的材料陈述的计算值属于正确记录。

	问题	rate	95% 区间
第一	数值在文本中	96.8%	94.9–98.0%
第一	文本涉及该材料	94.6%	92.3–96.3%
第一	both	93.4%	90.9–95.3%
首次	数值是一项测量	91.8%	89.1–93.9%
第二	数值在文本中	95.2%	93.0–96.8%
第二	文本中提到该材料的数值为	93.8%	91.3–95.6%
第二	both	91.8%	89.1–93.9%
第二	数值是一项测量	90.4%	87.5–92.7%

的化合物是二元或三元化合物。

各字段的覆盖情况并不均衡，覆盖率衡量的是抽取情况，而不是文献情况。图 3 展示了各项性质已填写的记录所占比例。 T_c 几乎处处都有，因为这是语料库中几乎每篇论文都会陈述的一项性质。 H_{c2}、相干长度、穿透深度和能隙只出现在少数记录中；而配对对称性和机制以文字形式报告，且

无法直接对其应用逐字核查，因此最为罕见。

IV. 审计幸存项

门控条件的效果取决于所提出的问题，而它们所问的是：论文是否针对该材料陈述了一个数值。它们并不判断论文是否正确。论文将某个数值陈述为计算结果时，这就是一条成立的记录：它不是错误，也不是观测值，而数据库会保留这条记录，并注明这一事实。处理流程自身历史中的三个例子展现了这种区别。单层 FeSe/$SrTiO_3$ 的转变温度 109 K 被记录到了块体 FeSe 名下，而块体 FeSe 的转变温度约为 8 K：该句确实写了 109 K，也确实提到了 FeSe，但含义并非记录所指的那样。MgB_2的 39 K 被记录到了 $YBa_2Cu_3O_{7-\delta}$名下，来源段落则同时给出了这两项信息。氢化物则以第一性原理研究中的数值录入数据库（LaH_{10} 在 288 K，YH_9 在 276 K）；另有一个 56 K 的压力点被同时写入四种化合物。

为了衡量这些内容有多少得以保留，我们对语料库进行分层抽样并审计两次：由两个未参与构建数据库的模型分别审阅每条记录所依据的整篇论文。样本共含 500 条记录，按材料族和转变—

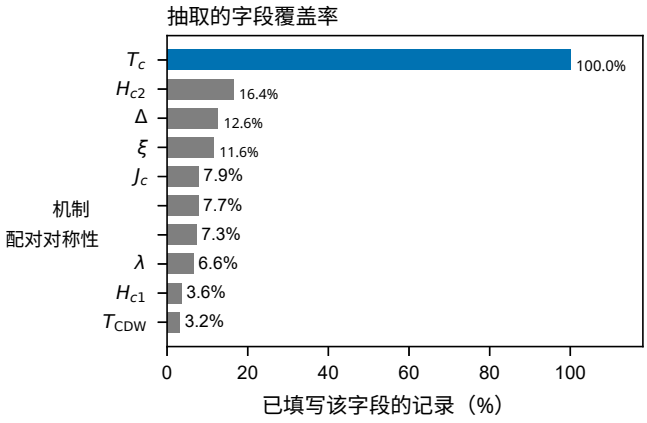

图 3. 抽取实际填写了多少字段。字段出现在 3,137 条记录中的比例。

温度区间分层抽取，分配比例与语料库一致，因此得出的比率反映的是数据库的比率，而不是过度代表稀有材料族的样本；每个非空的材料族和区间都包含在内，且抽样使用固定随机种子。处理流程已知其化合物存在问题的记录均作了标记，并单独报告：与其余记录相比，困难记录与第一位评审者的判定一致的频率较低（90.6 %，而其余记录为 94.5%）。

A. 双模型审计

该样本由两个未参与构建数据库的模型分别审计了两次。每个模型都拿到了数值所属的整篇论文，而不是数值附近的一段文本，并被问到相同的三个问题：文本中是否有这个数值，这段文本是否涉及该材料，以及这个数值是测量结果还是计算结果。第三个问题描述数值的获取方式，并不决定该记录是否成立：如果论文给出了针对正确材料的计算值，那么这就是一条正确记录。第一个模型是 DeepSeek-V4.1-Flash，第二个是 DeepSeek-V4-Pro。整篇论文对判断结果是否有意义至关重要：只看到数值所在句子的判定者无法判断后文是否提到了该材料，因此会拒绝本来正确的记录。

第一位评审者阅读的 500 条记录中，有 484 条（96.8%，95 % 威尔逊区间为 94.9–98.0%）在正文中明确给出了数值；其中 473 条（94.6%，92.3–96.3%）的句子讨论的是记录中所列的材料。这两个条件同时成立的记录有 467 条（93.4%，90.9–95.3%），本文其余部分采用的就是这一比率。难度较大的案例中，第一位评审者最不可靠：138 条此类记录中有 125 条（90.6%，84.5–94.4%）同时满足这两个条件，而其他 362 条中有 342 条（94.5 %，91.6–96.4%）；第二位评审者对它们的区分更明显，138 条中有 120 条（87.0%，80.3–91.6%），而其他 362 条中有 339 条

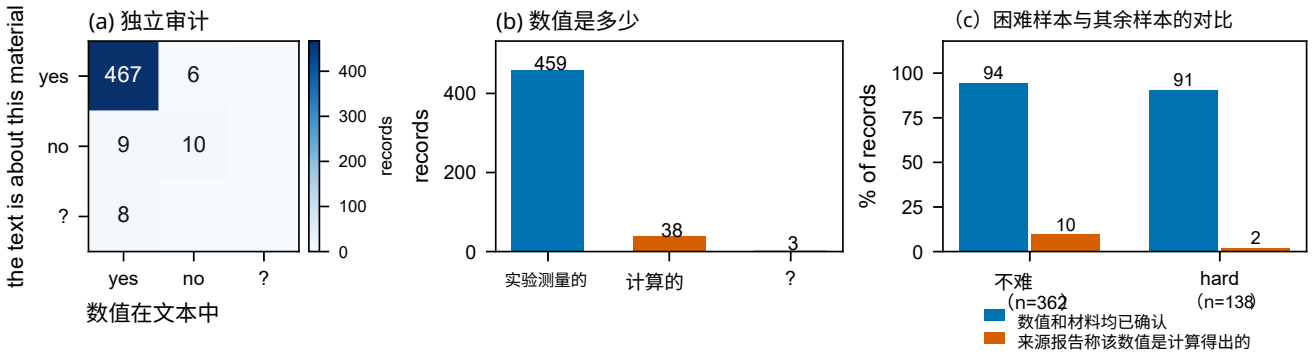

图4. 首次审计的发现。（a）文本中的数值与文本是否涉及该材料；第一位评审者在其判定的500条记录中确认了467条记录的这两项内容。（b）提取的数值类型：大多数是测量值，7.6%是来源论文报告为计算得出的数值。（c）将这两个比率分别按样本为构建困难样本而纳入的记录和其余记录进行区分。

其余 362 条（93.6%，90.6–95.7%）。

第一位评审者不接受的记录有 33 条。逐条人工检查后，它们分为两类。16 条记录的数值是评审者无法在正文中找到的：将某个家族的转变归在其某个成员名下（$BaFe_2As_2$ 在 29 K 时，论文给出的是“122”族的数值；层状氮化物 ZrNCl 在 26 K 时；Ce-115 族 $CeIrIn_5$ 在 2 K 时），从并非该化合物的样本中读出数值（来自一个结的 Nb 在 0.55 K 时；来自一个腔体的 Nb_3Sn 在 9 K 时），将增幅记作温度（MgB_2-GaP 在 0.8 K 时），将界限记作数值（CuH_2在 40 K 时，论文写的是“$T_c > 40$ K”），记录论文计算而非测量的数值（Cs_xMoS_2 在 5 K 时，其句子通过调节耦合常数确定了 T_c），以及陈述相反情况的内容（FeSe 在 6.5 K 时，句子称薄膜在降至 6.5 K 时仍未超导）。另外 17 条记录的数值虽在正文中，但评审者认为其所在句子并非在讨论记录中所列的材料；其中 7 条是评审者无法判断该句是否提到了这种材料的记录。

第三个问题询问这些数值是如何得到的。在审计的记录中，第一位评审者认为有 7.6%（5.6–10.3%）是来源论文报告为计算所得的数值（图4），第二位评审者给出的比例为 7.4%（5.4–10.0%）。这与语料库本身标记为计算所得的 5.9% 接近，而且两位评审者找出的是同一批记录：语料库也标记为计算所得的 23 个计算值，正是两位评审者各自找出的那 23 个；两位评审者都没有发现被语料库标记为计算所得的实测数值。逐条手动核查计算所得的记录证实了这一点：ScH_6 的记录值为 100 K，依据的句子写道：“我们使用 T_c 的典型取值 μ^*，例如 0.10 和 0.15，进行计算，得到约 100 K 的高临界温度”；$LiCaH_{12}$ 的记录值为 288 K，依据的是一份计算得到的氢化物列表以及对这些氢化物所能达到数值的预测。对于记录中所列化合物，论文计算并明确给出的数值，忠实记录了论文所述内容；而记录的标记则注明该数值是计算所得，而非实验测量的。数据库会保留这些数值并为其添加标记，以便读者可以只查询实测温度，也可以同时查询预测值。

第二位评审者与第一位评审者在 500 条记录中的 470 条上得出了相同结论：其中 453 条两者均接受，27 条两者均不接受，20 条则判定不同（表 II）。两位评审者意见不同时，分歧在于证据，而非数值。对于 70 K 下的 FeSe，第一位评审者接受该记录，第二位评审者则予以拒绝；对于 29 K 下的 $BaFe_2$砷$_2$，以及 10 K 下的 $FeSe_{0.83}$硫$_{0.17}$，第二位评审者接受了第一位评审者拒绝的记录；对于 41.9 K 下的 $(Li_{1-x}$铁$_x)$OHFeSe，以及 92 K 下的 YBa_2铜$_3$氧$_{7-\delta}$O，接受对方未接受记录的则是第一位评审者。仅前两个问题上，两位评审者就在三十条记录上意见不同；第三个问题上则有十三条。因此，两者之间的一致性是关于这两个模型的事实，并不能保证这些记录正确。

V. 分布顶部由什么构成

审计得出，语料库中计算值的占比为 7.6%，第二位评审者的判定为 7.4%——接近语料库本身所包含的 5.9%。如果能够定位这样的占比，使用起来会更加方便；而数据库包含了定位所需的信息：每条记录都保留一个标记，用于说明其来源是将转变温度报告为实验观测还是预测；同时，每条记录都保留了提取该数值的原句。

图 5 展示了该标记在整个语料库中的应用情况。在 3,137 条记录中，有 185 条（5.9%）被标记为非实验数据，但这些标记的分布并不均匀。在临界温度达到 25 K 及以上的记录中，9% 被标记；达到 50 K 的记录中，13% 被标记；达到 100 K 的记录中，21% 被标记；达到 150 K 的记录中，28% 被标记；达到 250 K 的记录中，41% 被标记。不同材料族之间相差一个数量级：在 158 条临界温度超过 100 K 的氢化物记录中，有 57 条（36%）被标记为预测值；在 125 条临界温度超过 100 K 的铜氧化物记录中，有 2 条（2%）被标记。

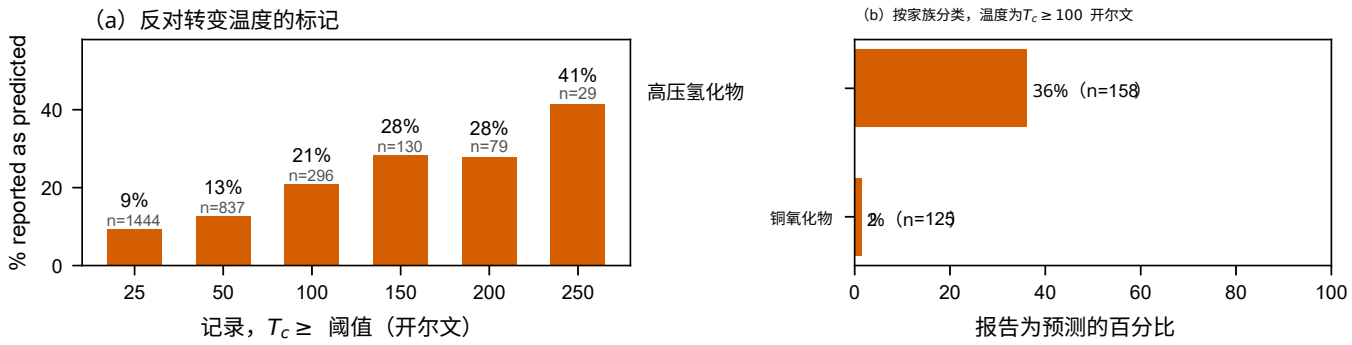

图5. 来源报告了预测的记录分布情况。(a) 在转变温度达到阈值的记录中，来源将转变温度报告为预测的记录所占比例。(b) 对于超过 100 K 的记录，按材料族统计同一比例；此类记录少于十条的材料族不予显示。该标记是抽取过程对来源论文的判读，并按记录而非数值设置。

该标记是抽取过程对其来源的自行解读，而审计则由一个未参与抽取的模型对同一批记录进行二次解读。在对 500 条经过审计的记录进行比较时，两种解读只在一个方向上存在分歧：在第一位评审者判定为计算所得的 38 条记录中，该标记也将其中 23 条标记为计算所得；而第一位评审者判定为实验测量的 459 条记录中，没有一条被该标记标为计算所得。因此，有 15 个计算值被存储为实验测量值。该标记得出的每个占比都是语料库中计算值占比的下界，而审计得出的 7.6% 是对语料库实际情况更准确的估计。

因此，数据库中最高的转变温度由两类陈述构成。排名最前的两条记录都是转变温度为 298 K 的氢化物，但类型各不相同：$LaSc_2H_{24}$，的句子记录了超导性已通过零电阻得到确认，且在外部磁场下 T_c 受到抑制；$Sr_2Sc_4H_{33}$，的句子则是来源于一篇预测结构论文的结构标签。超过 290 K 的五条记录中，有四条来自计算。

这是一个可以向语料库提出并得到回答的物理问题，也正因如此，这种形式的数据库值得保留。搜索高-T_c 超导体时，这些记录会最先出现；而寻找可制备材料的读者与寻找目标预测的读者，所需要的记录各不相同。这里的区别按记录标注，任何读者都可以对照数值所依据的句子进行核查。

VI. 数据库揭示了什么 S

数据库不仅记录了文献，也是一个语料库，可借此考察该领域中的各种关系。图 6 从保留记录中绘制了其中四种关系。这些图表由项目网站绘制；这里用相同的数据重新绘制，以便印刷。

前两项是任何超导体汇编都必须通过的检查项，而数据库像文献中的情况一样通过了这些检查。

上临界场会随转变温度升高，这一点与文献中的情况一致。在语料库涵盖的三个数量级范围内，T_c 二者呈现出相关性，这正是McMillan公式的基础；氢化物位于这一范围的高T_c端，上临界场最高达158 T。不过，语料库中上临界场最高的并非氢化物：一种铜氧化物在87 K时达到600 T，铁基化合物则达到377 T（图6a）。能隙在三个数量级的范围内也随转变温度变化，而这些记录分布在弱耦合线两侧：76%位于其上方，远低于该线的记录大多来自转变温度较低的材料（图6b）。这些散点中偏离较大的记录，问题并不在于能隙本身：提取列混合了不同的物理量，少数记录所标注的数值实际上属于其他物理量。

第三个面板展示了数据库自然生成的排名，也是本次审计关注的排名。各自家族中排名居前的两条记录，正是本文此前已指出的错误：铜氧化物的 175 K 这一数值是“$T_c = 0.15$ 对应约 175 K”，这是一个被存储为转变温度的标度结果；铁基的 109 K 则属于 FeSe单层，却被记录为块体 FeSe。第三条记录是 MgB_2 在 45 K 的记录——MgB_4P_4，是该家族中数值最高的记录——抽取过程本身已将其标记为计算值。若读者照单全收这份排名，就会把一种铜氧化物排在所有实验测量的铜氧化物之上。氢化物族中排名第一的记录为 298 K，其对应的句子报告了一次实验测量。

第四个面板展示了这个领域原本需要逐篇阅读论文才能得到的内容：镍酸盐的压力依赖性、十一条曲线、$La_{3-x}Nd_xNi_2O_7$ 在 6 至 45 吉帕之间的穹顶曲线、延伸至 70 和 80 吉帕的两种三层化合物，以及其中最高的转变温度——93 开尔文（图 6d）。这些曲线正是 SuperCon 配套数据集的贡献，也是最能体现语料库价值的例子：它保留了测量中的每个数据点，而不是每种化合物只保留一个数值。

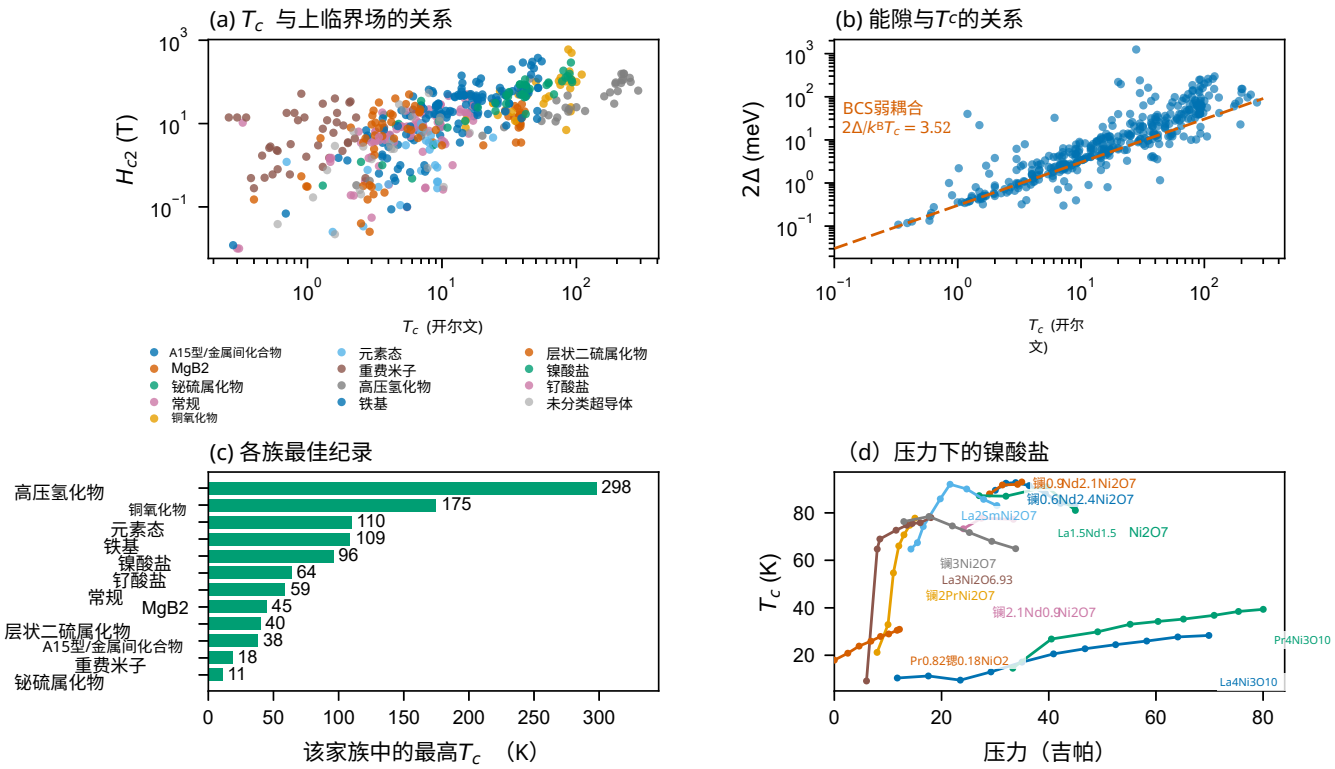

图6。四种关系。（a）–（c）根据保留记录绘制：（a）对数坐标轴上转变温度与上临界场的关系，按家族着色。（b）能隙与转变温度的关系，并以BCS关系的弱耦合极限 [13]，$2\Delta/k_{\mathrm{B}}T_c = 3.52$作为参考。（c）每个家族中记录的最高转变温度。（d）镍酸盐的超导转变温度对压力的依赖关系。

VII. 方法

A. 语料库和文档

论文通过无人值守下载循环从出版商处收集，该循环按期刊、作者和材料名称检索文献，并对已获取的 PDF 文件建立内容哈希缓存。文档使用 docling 转换为文本 [12]，后续每个步骤读取的都是转换生成的 Markdown。无法转换的论文，以及文本中完全没有超导量的论文，都不会提供任何内容。

B. 抽取

文本通过一个针对组分及其是否经实验验证的提示词，以及七个特定维度提示词交给大语言模型（MiniMax-M3.1-Flash）：基本超导性质、压力依赖性、同位素效应和 Uemura 分析、比热、费米参数和电阻率、晶体结构，以及制备。七个提示词中，有三个只会在论文涉及相应主题时提出——例如，文本提到同位素时才询问同位素效应，提到费米面时才询问费米面——而另外四个则始终会提出，因此，对于不涉及同位素效应的论文，不会为此项付费。答案会按组分合并，每种组分对应一条记录，并以组分本身作为键，因此，一个提示词找到的性质与另一个提示词找到的性质最终会出现在同一行中。

抽取语料库共调用模型 115,435 次，消耗 7.3×10^8 个词元，其中 1.0×10^8 个为补全词元，而其中 30% 为推理词元。这里的成本指阅读所耗费的成本；检查阅读结果的成本，则是本文其余部分所衡量的内容。

C. 门控条件

对合并记录应用四项检查，并对整个数据库应用第五项检查。

逐字证据。 每个数值都必须附带论文中出现的引文。引文和论文在比较前会经过相同的规范化处理：兼容性折叠（这很重要，因为转换会将数学内容呈现为 Unicode 斜体，而模型会写成普通字母）、连字展开、大小写折叠、将三种破折号统一折叠为一种、重新连接被换行符拆分的单词，以及

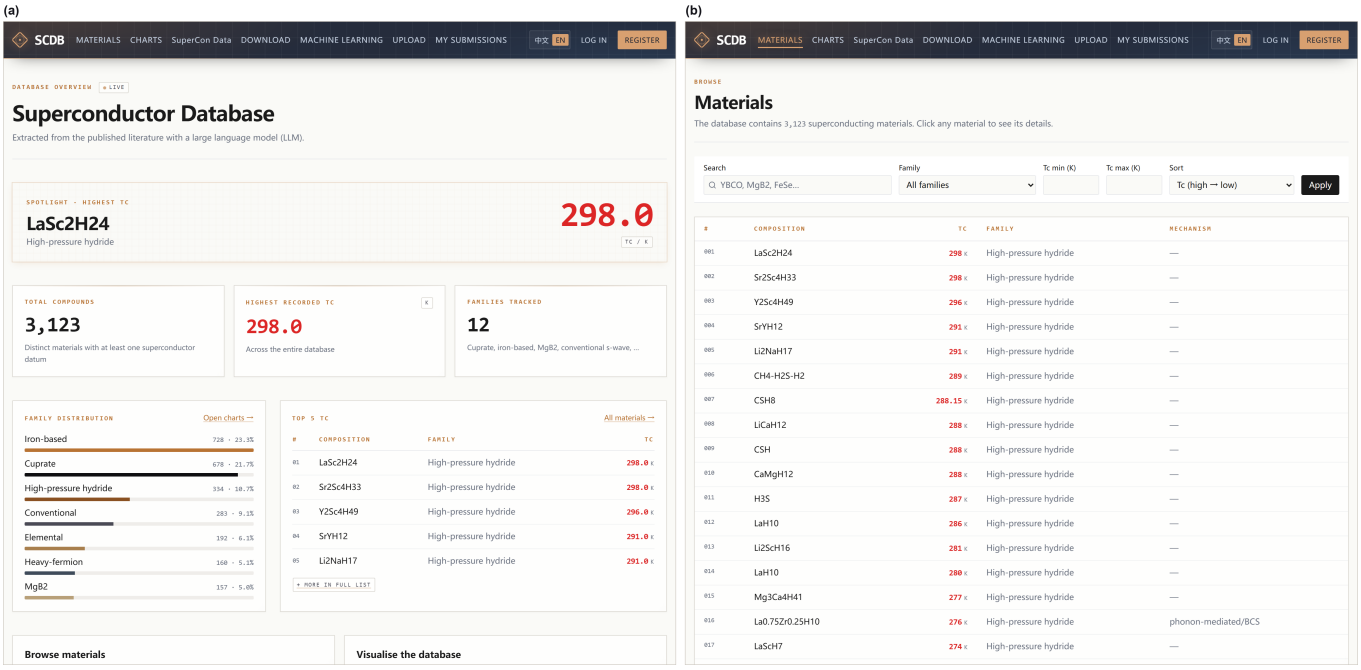

图7. 数据库作为网站。(a) 主页，展示语料库的规模、最高记录以及其中包含的材料族。(b) 材料页面，列出记录，并可按组分、家族、转变温度范围和排序方式筛选。截图为英文版；该网站也提供中文版。

标点符号和空白字符的合并。仅凭子字符串匹配，不接受短于 24 个字符的引文，因为“NA”或“临界温度 200 开尔文”会匹配任何碰巧印有该数值的论文；第二轮会在移除所有空白字符后比较二者，从而捕获转换过程中被换行拆开的词元。两个非数值标签——配对对称性和机制——也遵循同一规则，因为它们同样是关于化合物的声明。被接受的引文会与数值一同存储。

材料绑定。 数值必须出现在指明该材料的句子中，可以直接指明、通过别名指明，或由上下文指明。这样就能防止把“$YBa_2Cu_3O_7$= 39 K”（这是 MgB_2 的数值）收入数据库。

数值的获取方式。 通过阅读包含该数值的句子措辞，判断这个数值是实验测量的还是计算得出的。预测性的措辞会使数值被排除，观测性的措辞则会使其保留。系统会记录判定理由，而不是更正数值，以便依据来源审计这一决定。如果措辞无法得出结论，就保留该数值，并在记录中保留抽取模型对数值性质的判定；数据库标记记录的是这一判定，而不是措辞。

物理合理性。 超出其所属家族可达到范围的转变温度将被拒绝：除氢化物外，其他所有材料的上限为 200 K；氢化物的转变温度可以更高。对于单元素超导体，则采用更严格的界限，因为它们的转变温度已足够明确，可以识别明显不匹配的情况。该检查有意保持保守，因为拒绝一条真实的测量记录会造成损失，而接受一条归属错误的记录则会在不知不觉中破坏数据库。

跨论文一致性。 一篇论文报道的化合物若也见于其他论文，就会有多个数值；仅与单篇论文核对，无法发现论文自身报告的测量值虽出自该论文、却不适用于该材料的情况。彼此相差不超过 1 K 的数值会归并到同一簇；若没有任何簇占主导，则保留文档记录更完善的记录。随后，将描述同一化合物且具有相同转变温度的行合并为一行，并纳入它们各自包含的所有字段。若合并后的行对数值的获取方式存在分歧，则保留来源报告为观测的记录。对于语料库在其他地方已有测量值的化合物，其计算值会被剔除：该测量值已是数据库中该化合物的条目，而计算值只是对它的另一种、通常更高的表述。应用这两条规则后，记录数从去重后的 3,185 条降至 3,137 条；其中 185 条被标记为计算值，而按同一规则原有 233 条，按此前规则则有 261 条。

D. 审计

审计样本从语料库中抽取，按材料族和转变温度区间进行分层，并按各单元在语料库中的占比进行分配，因此审计得出的比率反映的是数据库的比率，而不是过度抽取稀有材料族所得的比率：铁基材料占样本的 23.0%

，而在语料库中占 23.4%；铜氧化物占 21.2%，而在语料库中占 21.7%；氢化物占 10.4%，而在语料库中占 10.6%。所有非空单元都包含在样本中，处理流程已知难以处理的组成会被标记并单独报告；抽样过程采用固定随机种子，因此可以再次抽取同一样本。抽取的 500 条记录仍全部存在于语料库中，因此评审者判定的比率就是数据库本身的比率。两位评审者各收到 500 条记录。

第一位评审者是 DeepSeek-V4.1-Flash，通过其自身的应用程序接口调用；它与执行抽取的模型属于不同家族。提供给它的是组分、已存储的数值，以及数值出处论文的全文——也就是已发表 PDF 的转换文本，而抽取过程读取的正是这份文本——随后它回答三个问题：文本中是否出现了该数值、文本讨论的是否是这种材料，以及该数值是测量结果还是计算结果。第二位评审者是 DeepSeek-V4-Pro，即同一提供商推出的规模更大的模型；它拿到相同的论文，并被要求回答同样的三个问题。答案的意义取决于提供整篇论文：只看到数值出处句子的评审者无法判断该材料是否在后文才被点名，因此会错误地否决正确的记录。评审者看到的仍是转换后的文本，而非排版后的论文；这限制了此类判定结果所能证明的内容，相关问题将在讨论部分进一步探讨。

按照同一原则，在上述规则写定之前，我们手动删除了一条记录：其中的句子称转变温度可能达到300 K，这是一个预测值，而非数值；若将其作为数值存储，就会使一种铜氧化物的转变温度高于所有实验测量的铜氧化物。

比率附有95% Wilson评分区间，在样本量较小的情况下，这种区间也能合理反映情况。

E. 软件

抽取、门控条件、合并和审计均使用 Python 完成；图表则使用 matplotlib，根据冻结的语料库快照绘制。本文引用的每个数字，都是由读取相关文件的脚本生成的，而非手工得出。

VIII. 数据库的实际使用

数据库是公开的，其背后的机制也同样公开：一个网站（http://isspfml.cn）按家族和转变温度提供材料，并可将语料库中的材料和 SuperCon 数据集导出为 CSV、JSON 或 JSON Lines 格式，每个家族一个文件。相关机制也都开放：提取处理流程、门控条件和审计脚本均可在 https://github.com/isspwjlu/SCDB获取。该网站也提供了一个入口：读者可以提交语料库中尚未收录的论文 PDF，经审核者审阅后，从中读取的记录便会加入语料库。SuperCon 配套数据集 [2]收录了 NIMS 数据库经整理的转变温度，其中镍酸盐部分来自我们的数据库。

因此，该网站提供 15,408 条记录，其中 15,223 条（98.8%）为实验测量的转变温度，185 条（1.2%）是其来源报告的计算值；每一行都会注明所属类别。两种占比必须分开解读，因为它们来自两个不同的语料库。该网站的实验测量记录占比很高，是因为其提供的内容中，超过一半来自人工整理的 NIMS目录，其中每一行都是一项测量，无需设置标记；本文抽取的 3,137 条记录中，94.1% 为实验测量的，而审计显示，其中计算值占 7.6%。

该语料库也是训练数据。根据化学式进行预测 T_c 是一项由来已久的任务 [14, 15]；人们已采用多种机器学习方法来完成这项任务，其中包括梯度提升和随机森林[16–20]，。该团队沿着相同思路开发的 SuperconPred工具包，会将化学式转换为 45 个特征，并使用随机森林、梯度提升或高斯过程回归进行预测 [21]。用于这一目的的审计数据库，其意义不仅在于规模大于人工整理的数据库，还在于可以追溯基于该数据库训练的模型所用的每个训练数值的来源。如果模型从一条错误归属的数据中学习 T_c，它学到的就是论文中并不存在的化学事实；如果它从计算得出的数值中学习 T_c ，则可以证明它学到的是预测，而非观测。

这些门控条件并非专为该语料库设计：五项条件都是针对句子和材料编写的，而非针对模式中的某个字段，因此同一组条件也可用于其他语料库，或其他团队整理的数据库。无法直接套用的是结果：这里的比率反映的是该语料库的情况；若要审计另一个语料库，就必须重新进行审计。

IX. 讨论

门控条件可以通过阅读文本排除那些能够排除的问题。挖掘文献以寻找材料的方法一直要求结构化输出、模型与人工之间的一致性，以及每个数值的来源记录 [5, 22–25]；这篇论文提出的问题则不同：筛选后留下的数值中有多少是正确的？对于这个语料库，答案是 93.4% 的经审计记录同时满足这两个条件。审计接着衡量答案的另一面：留下的数值中有多少是预测值而非观测值？在该样本中，这一比例为 7.6%。门控条件无法回答论文自身提出的问题，即某个数值是如何得出的，因为这并非阅读错误：论文正确地陈述了该数值，所指材料也正确，并且只说这个数值是计算得出的。这样的数值准确记录了论文所述内容，但它并非实验测量结果。

这些数值大多出现在氢化物研究中，其中预测的转变温度达到 200 开尔文区间，并且与实验测量的转变温度出现在同一段落中。这两类数值都应纳入语料库；数据库应向读者说明哪些是哪些，并保留能够判定它们性质的那句话。

改进的方向不是反复审视同一句话，而是改变对它的要求。目前，核查数值获取方式的检查会查找数字附近的措辞；如果一句话同时包含两类措辞，它就会优先判定为观测：例如，一句话将“计算得到的 T_c 数值”与“实验测量的数值”并列，就会被标记为实验测量的；另一句话写道“我们报告的 T_c 在两个相中几乎相同，其峰值达到 34.6 K”，也会被如此标记。漏掉的 15 条记录中，有两条就是这样存储的。如果论文在预测结构表格中给出数值，检查也会失效，因为那里没有可供查找的措辞。若检查要求句子明确说明数值是如何获得的（实验测量还是预测），并让表示预测的措辞优先于表示观测的措辞，就能将 SnH_{12} 这一行和 YH_6 这一范围正确分类，也能识别出那条被存储成单个数值的范围。

审计本身有三个局限，应当说明。两位评审者都是模型，而非读者：它们看到的是论文转换后的文本，而不是论文原文。它们拿到的是整篇论文，但在 500 条记录中仍有 20 条意见不一致——因此，它们之间的一致性只是关于这两个模型的事实，并不能保证这些记录正确。样本也只是样本：它按语料库的比例抽取，因此其比率在抽样误差范围内反映语料库的比率，而此处引用的每个比率都带有该误差的区间。

对同一批记录的两种判读在总体水平上相符，这正是按语料库比例抽取样本的目的：标记将语料库中 5.9% 的记录标为非实验数据，而审计重新阅读相同句子后，得出的比例为 7.6%。两者并非完全一致；差异正是只根据对来源论文如何表述数值的一次判读来设定标记所付出的代价：它捕捉到评审者发现的 38 个计算值中的 23 个，漏掉其余 15 个，但没有把任何记录错误地标成相反类别。因此，语料库中的 5.9% 是下界，审计得出的 7.6% 是更好的估计，真实占比则介于两者之间。两种判读完全一致的是这些记录的身份：标记为计算值的 23 条记录，正是评审者标为计算值的那 23 条。因此，想要获取计算值的读者可以按该标记筛选，并准确找出相应记录。两种判读还都将它们归入高-T_c 尾部和氢化物之中，而这正是两类数值之间的区别最为重要的地方。

第二个限制因素是语料库本身。论文仅限于能够获取的那些，而该领域的数据库则未能纳入：NIMS SuperCon 数据集 [2],Yamaguchi 及其同事整理的超导体语料库 [10], Foppiano 的自动抽取工作及其同事 [9], 以及最近汇编的超导体及其晶体结构数据集 [11]。本研究的语料库是对这些资源的补充，而非替代：它规模较小，而且其中每条记录都可以对照某个句子进行核查。

与先前的工作相比，差异在于性质不同。利用大语言模型构建的数据库，即便有所核查，也只是让另一个模型阅读相同的句子来进行核查 [7], 而从文献中抽取超导体则被当作构建语料库的问题来处理 [9, 10]而不是追问抽取后哪些信息得以保留。将句子与数值一同存储，就让这个问题变得可以回答；这里的答案是，这种方式构建的数据库可用，而且之所以可用，是因为包含计算值的记录都标明了这一点。标记这些记录成本很低：只需在记录中添加一个字段，并保留数值所源自的句子，以便对照核查。

数值是如何得出的，这个问题并非超导体所独有，处理方式也并非如此：语言模型的幻觉如今已有编目 [26], 负责阅读材料文献的智能体也被赋予了同类任务 [27, 28], 而让人在抽取过程中参与的工具也正在开发 [29]。超导性案例的特别之处在于，它涉及一个既容易核查、又容易出错的量。氢化物是最典型的例子。它们的预测转变温度高达 200 K 左右 [30, 31], 与实验测量的转变温度出现在同一段落中，而且搜索高温T_c 超导体时，它们会最先出现在结果中。

有两点使这成为一个物理学期刊的问题，而不是数据处理问题。第一点是答案所涉及的内容：预测的转变温度和实验测量的转变温度回答的是不同的问题；语料库若不说明二者分别是什么，就无法回答其中任何一个问题；氢化物中这两类温度最为接近。第二点是，答案关乎文献本身，而非某一个语料库：门控条件和审计是针对句子制定的，无需改变测量内容，就可以应用于任何一组超导论文。

数据可用性

材料记录、每个数值所摘自的句子、来源 DOI 以及派生字段，可从项目网站（http://isspfml.cn）获取，也可获取上述导出文件，格式包括 CSV、JSON 和 JSON Lines。本文使用的语料库快照（26,210 篇论文、3,137 条材料记录，涵盖 1,371 种成分）是 2026年11月21日发布的版本。

代码可用性

提取流程、门控条件、整合流程，以及读取审计判定结果的脚本，均已发布在 https://github.com/isspwjlu/SCDB。数据集不在该代码库中：NIMS 为研究用途分发 SuperCon 目录，而材料记录则是网站所提供的数据库。因此，审计样本和两位评审者的判定结果也不在其中；不过，用于处理这些数据的脚本已发布，因此数据到手后，便可使用相同文件重新计算本文引用的数值。

作者贡献

W.J.L. 设计了抽取流程和验证方案，编写了代码，生成了统计数据和图表，并撰写了本文。

人工智能使用声明

语料库由 MiniMax-M3.1-Flash 抽取，提取记录的独立审计则使用 DeepSeek-V4.1-Flash 和 DeepSeek-V4- Pro 完成；这些模型仅用于文中所述用途，没有用于其他任何用途。Claude Code 用于编写抽取、验证和绘图代码，生成本文中的图表，以及起草和修订稿件；此处报告的数值由已发布的脚本计算，其解读由作者负责。

致谢

作者感谢 ClaudeCode 负责编程、图表、稿件撰写和网站制作，并感谢 docling 的开发者 [12] 提供文档转换功能，而处理流程依赖于此。

利益冲突

作者声明不存在利益冲突。

附录A：作为网站的数据库

这个数据库不仅仅是一张表；它还是一个网站，让读者无需编写代码即可使用，也是大多数读者接触本文所述材料的形式。论文。http://isspfml.cn。图7 展示了承载该数据库的两个页面。

材料 页面（图 7b）是入口页面。记录按转变温度、家族和机制列出，并可通过搜索组分、按家族筛选、设定转变温度范围以及排序来缩小范围。菜单中的家族是抽取流程实际识别出的家族，并显示了各自的数量，因此，想查看镍酸盐的读者可以先了解其数量，再进行查询。点击一条记录即可打开详情页：页面顶部显示组分和转变温度，以及家族和来源论文的数字对象标识符、该行记录的来源；随后按抽取流程填充的分组列出性质——基本超导量、压力依赖性、同位素效应、比热、费米参数、晶体结构和制备条件。若论文未说明某项性质，该项就留空，而不填入默认值；门控条件正是按此规则处理数值。

主页（图7a）给出了语料库的规模、最高记录及其包含的材料族，并在下方列出其中最高的五个转变温度。

图表页面展示了转变温度的分布和各材料族的组分，并在旁边列出图6 中的四种关系。本文的图表就是重新绘制的这些图表。

下载页面是导出的途径。语料库材料和 SuperCon 配套数据集以 CSV、JSON 或 JSON Lines 格式提供，可按每个材料族单独提供文件，也可将多个材料族打包成一个档案；每份文件都附有提取每个数值的原句，因此，读者若想对照来源核验某个数值，无需编写任何代码即可完成。页面对公众开放；导出则需要注册账户，以便统计下载次数。

上传页面是提交内容的入口。注册读者可以提交论文的 PDF，文件会在存放语料库的机器上读取，使用与读取语料库相同的处理流程，并遵循相同的门控条件：数值必须出现在文本中，必须属于该材料，而且必须记录其获取方式。抽取结果会展示给审核者——包括组成、转变温度，以及提取每个数值的原句；审核者批准的内容将并入语料库，同时注明其来源论文，并标记为一项贡献。通过这种方式提交的记录，须符合与语料库扫描所得记录相同的标准；这一路径的数量与另一条路径分别统计，因为两者属于不同的群体。本文描述的语料库是2026年11月21日发布的版本，早于这一路径的推出，因此目前尚无记录通过此途径收录。

该网站已上线，抽取仍在继续，因此其显示的计数会不断增加；图7 是在本文所述版本发布时制作的。

[1] H. K. Onnes, 《超导性的发现》（1911年）。[2]日本国立材料科学研究所, 《SuperCon数据表》, 由日本国立材料科学研究所（NIMS）公开的超导材料数值数据表（2022年）。[3] S. Grazulis、A. Daskevic、A. Merkys、D. Chateigner、L. Lutterotti、M. Quiros、N. R. Serebryanaya、P. Moeck、R. T. Downs 和 A. L. Bail, 《Nucleic Acids Res.》40，D420（2011年）。[4] A. Jain、S. P. Ong、G. Hautier、W. Chen、W. D. Richards、S. Dacek、S. Cholia、D. Gunter、D. Skinner、G. Ceder 和 K. A. Persson, 《APL Mater.》1，011002（2013年）。[5] E. Kim、K. Huang、A. Saunders、A. McCallum、G. Ceder 和 E. Olivetti, 《Chem. Mater.》29，9436（2017年）。[6] Y. Zhang、S. Itani、K. Khanal、E. Okyere、G. Smith、K. Takahashi 和 J. Zang, 《J. Magn. Magn. Mater.》597，172001（2024年）。[7] S. Itani、Y. Zhang 和 J. Zang, 《《自然·通讯》》16，9415（2025年）。[8] S. Itani、Y. Zhang 和 J. Zang, 《《计算材料科学》》253，113855（2025年）。[9] L. Foppiano、P. B. Castro、P. O. Suarez、K. Terashima、Y. Takano 和 M. Ishii, 《Science and Technology of Advanced Materials: Methods》3，2153633（2023年）。[10] K. Yamaguchi、R. Asahi 和 Y. Sasaki, 《Adv. Eng. Inf.》54，101768（2022年）。[11] T. Sommer、R. Willa、J. Schmalian 和 P. Friederich, 《Sci. Data》10，816（2023年）。[12] P. Staar, 载于*Proceedings of the 2020 ACM Symposium on Document Engineering* 2026 1 1 （ACM, ）第–页。[13] J. Bardeen、L. N. Cooper 和 J. R. Schrieffer, 《超导理论》（1957年）。[14] K. Hamidieh, 《《计算材料科学》》154，346（2018年）。[15] V. Stanev、C. Oses、A. G. Kusne、E. Rodriguez、J. Paglione、S. Curtarolo 和 I. Takeuchi, 《npj Comput. Mater.》4，29（2018年）。[16] A. D. Smith、S. B. Harris、R. P. Camata、D. Yan 和 C. C. Chen, 《物理评论B》108，174514（2023）。[17] S. G. Jung、G. Jung 和 J. M. Cole, 《化学信息与建模杂志》64，7349（2024）。[18] H. Gashmard、H. Shakeripour 和 M. Alaei, 《科学报告》14，3965（2024）。[19] Y. Zhang 和 X. Xu, 《Physica C：超导及其应用》595，1354031（2022年）。[20] N. Sizochenko 和 M. Hofmann, 《分子》26，8（2020）。[21] X. Li、W. Tu、R. Lv、L. Liu、D. Shao、Y. Sun 和 W. Lu, 《物理评论B》113，054521（2026年）。[22] M. P. Polak 和 D. Morgan, 《自然·通讯》15，1569（2024年）。[23] M. Schilling-Wilhelmi、M. Rios-Garcia、S. Shabih、M. V. Gil、S. Miret、C. T. Koch、J. A. Marquez 和 K. M. Jablonka, 《化学学会评论》54，1125（2025年）。[24] G. Lei、R. Docherty 和 S. J. Cooper, 《数字发现》 3，1257 2024( )。[25] V. T. Phi 和 Y. Matsumoto, 载于 *Findings of the Association for Computational Linguistics: EMNLP 2025* （计算语言学协会，2025年），第 2369–2388 页。[26] A. Alansari 和 H. Luqman, 《计算机科学评论》61，100970（2026年）。[27] A. Chaudhari、J. Ock 和 A. B. Farimani, 《通讯·材料》7，131（2026年）。[28] S. O. Alfred 和 V. Sundararaghavan, 《计算材料科学》272，114862（2026年）。[29] X. Wang、S. L. Huey、R. Sheng、S. Mehta 和 F. Wang, 《坎贝尔系统综述：更好的证据，创造更美好的世界》21，cl2.70073（2025年）。[30] A. P. Drozdov、M. I. Eremets、I. A. Troyan、V. Kseno-fontov 和 S. I. Shylin, 《自然》525，73（2015）。[31] A. P. Drozdov、P. P. Kong、V. S. Minkov、S. P. Besedin、M. A. Kuzovnikov、S. Mozaffari、L. Balicas、F. F. Bal-akirev、D. E. Graf、V. B. Prakapenka、E. Greenberg、D. A. Knyazev、M. Tkacz 和 M. I. Eremets, 《自然》569，528（2019）。