arXiv:2611.28338v1 [cs.DL] 20 Nov 2026

MagTrace：追踪不断演进的科学文献中的磁序

Mengnan Sun*[1,2] 和 Cheng Peng†[1]

[1]斯坦福材料与能源科学研究所，SLAC国家加速器实验室，2575 Sand Hill Road，门洛帕克，CA 94025，美国[2]斯坦福大学，斯坦福，CA 94305，美国

摘要

许多磁性材料研究都从其磁序入手，但人工查阅论文来确认磁序可能需要数小时，而精选数据库也可能不完整且过时。在这项工作中，我们构建了 MagTrace，这是一种语言模型智能体，只需输入材料名称，便可搜索四个文献数据库、阅读开放获取全文，并报告当前文献所支持的磁序；每种磁序都附有置信度分数和可供读者核查的引文。我们针对18种材料评估其性能，并将结果与根据已发表论文编制的答案键进行比较，同时考察了两种检索程度不同的条件。采用0到1的分数衡量正确性、细节、置信度最高的磁序是否为电流所支持，以及引文完整性后，MagTrace得分为0.839；同一底层语言模型在获得一次文献检索所得的摘要时得分为0.676，而凭记忆作答时得分为0.548。仅进行一次搜索时，模型能够找出正确类型的磁序，但在85%的结果中遗漏了传播矢量；传播矢量描述图案如何在材料的晶体结构中重复。从记忆作答时，模型经常引用一些标题，却找不到对应的已索引论文（在我们能够核查的磁序中占25%）；而对于确实存在且有摘要的被引论文，其中17%的摘要与主张相矛盾，相比之下，MagTrace的这一比例为7%。

1 引言

磁序描述了材料中磁矩的指向，以及该图案如何在其内部晶体结构——原子的规则排列——中重复。磁序主要通过中子衍射进行测量：由于中子自身带有磁矩，它们会与材料中的磁矩发生散射，而产生的散射图案则显示出这些磁矩的排列方式 [1]。由于磁序会影响材料的电子结构、对称性和宏观性质 [2, 3], ，准确且最新的磁序知识对于解读实验、预测材料行为，以及发现新材料或现有材料的新用途都至关重要 [4]。例如，对一种磁性

材料进行第一性原理计算时，必须从初始磁序开始，而由哪种磁序作为起点十分重要：如果初始磁序选择错误，即使计算本身正确，也可能得出错误的结论。

材料的磁序记录为有序类型（例如，铁磁性、反铁磁、亚铁磁性、非共线）与传播矢量$\mathbf{k}$的组合；传播矢量描述了图案如何从一个晶体学晶胞重复到下一个晶胞。在数据允许的情况下，描述材料磁序时还可以包含对称性标签，例如磁（舒布尼科夫）空间群或晶体空间群的不可约表示 [5]。本文中，*磁序* 指这一整体组合，而*有序类型* 仅指有序类型本身。

*通讯作者：msun29@stanford.edu
†通讯作者：cpeng18@stanford.edu

如今，磁序也用作磁性机器学习模型训练数据集中的标签，而这类数据集需要涵盖数千种材料的磁序。对材料科学家而言，核查单个材料的磁序意味着人工查找所有相关论文，弄清每篇论文描述的磁序，并在研究结果不一致时判断应当相信哪一项。每种材料的核查过程可能需要数小时至数天；如果完全依靠人工数据收集，就难以获得机器学习所需规模的数据集。因此，这类数据集中的标签往往来自计算或整理数据库，而非通过查阅文献进行人工核查 [6]。

因此，精选数据库变得非常重要，因为科学家会查询这些数据库，以获取所需的材料信息。然而，精心整理的数据库中的每个条目都反映了创建该条目时的文献状况，这可能带来两个主要问题。首先，数据库并不完整：新的材料仍在不断被报道，而且大多数磁性材料的磁序都未存储在任何数据库中。截至2026年10月，MAGNDATA [7, 8], 这一测量磁序的主要精选数据库收录了2,447个条目，而由语言模型根据已发表论文构建的东北材料数据库列出了67,573种磁性材料，但至多只记录每种材料的磁序类型，而没有记录其磁序的完整描述 [9]。在我们评估集中用于代表广泛材料类别的18种材料中，只有8种在MAGNDATA中有条目。其次，数据库可能已经过时：随着新的实验依据出现，已有的磁序可能会被修订。UO_2 就是一个例子：经过数十年的实验，其磁序被修订过 [10]。因此，同一种材料往往会有多个条目。在MAGNDATA中，1,558种条目提供了纯化学式的材料中，有464种有两个或更多条目（附录A）。其中一些条目描述不同相或不同条件，另一些则是不同论文对同一相的报道，有时相隔数十年。在我们的评估集中，有magndata 条目的材料中，有3种材料的条目并不完全一致（§4.2）。找到多个条目的科学家必须回查论文，才能了解哪一个描述的是当前认知下的材料。

先前研究曾尝试通过三种方式，以更大规模提供磁性数据。第一种方式计算磁序，而不是收集测量得到的磁序。例如，Horton等人 [11] 构建了一种高通量工作流，用于预测每种材料的基态共线磁序。计算得到的磁序是一种预测，而非准确的测量结果，并且该工作流仅限于共线磁序。计算结果之间也可能存在分歧：当对三个大型材料性质计算数据库进行两两比较时，对于它们共有的材料，其中最多有15%的材料是否具有磁性这一点存在分歧 [12]。第二种方式是自动从论文中提取数据。文本挖掘工具和语言模型从研究论文中提取性质值 [13, 14], ；对于磁性材料，它们已被用于构建相变温度数据库 [15, 16], ，其中最大的是东北材料数据库。然而，这些数据库除了磁序类别之外没有更多信息。第三种方式是在经过整理的记录本身上训练机器学习模型。例如，Fahmy [17] 在magndata条目上训练分类器，以预测材料的传播矢量是否为零，并利用这些分类器标记出超过6,840条可能被误标为铁磁性的计算条目。这类模型从数据库中学习，并将其中的条目视为正确，因此也会继承数据库中的缺失信息和过时条目。因此，这些方法都无法取代经过整理的实测磁序记录，也没有一种方法会将该记录与当前文献进行核对。

直接询问语言模型某种材料的磁序，同样不是可靠的替代方案。先前研究发现，语言模型可能会引用并不存在的论文 [18]。尽管人们认为较新的模型已有大幅改进，但其可靠性仍存疑。能够阅读科学文献的语言模型智能体是一种较新的选择。据我们所知，目前没有智能体会通过阅读文献来报告磁序，因此我们将自己的智能体与最接近的通用智能体进行比较；这些智能体与我们的智能体承担相同的任务：搜索文献、阅读论文，并为每个答案提供引文。比较结果显示，这些智能体已经能够解决我们任务中的哪些部分，又有哪些部分尚未解决。例如，PaperQA 及其后续版本 PaperQA2 会搜索论文全文，并为科学问题提供带引文的答案 [19, 20], ；PaperQA2 还会标记论文之间的矛盾。OpenScholar 对 4500 万篇开放获取论文也能完成同样的工作，其引文准确度与人类专家撰写的引文不相上下 [21]。化学和材料科学领域的类似智能体会运行工具、开展实验和模拟 [22–24], ，或者像 Ansari 和 Moosavi 的智能体那样[25], ，将论文中报告的数值收集到数据集中。与我们的用途最接近的是 LitMOF：它会将晶体结构数据库中的条目与其来源论文进行核对，并修正错误条目 [26]。然而，这些智能体要么回答自由文本问题，要么收集数值而不对其进行判别，要么将数据库条目与其自身的来源论文进行核对。它们都没有被要求给出足够详细、可与论文核验的磁序，也没有被要求判断多个已发表结果中哪一个是当前结果。

表1： 与本研究并列展示以往方法如何提供材料的磁序。前五行是针对磁性数据的方法。接下来两行是为与我们类似的目的读取论文的智能体，但并非用于处理磁序。

方法	测得的磁序	完整磁序（类型和 k）	电流文献	选择当前排序
整理数据库 magndata [7, 8]	yes	yes	不，固定为条目	no
计算阶数 [11]	no	仅共线	no	no
文本挖掘数据库 [9, 15, 16]	测量数据，而非磁序	不，只有种类和温度	否，固定为提取	no
在 magndata 上训练的分类器以及计算标签 [6, 17]	从……中学到数据库条目	不，是一个预测的类别	no	no
LLaMP [29]	no	不，只有 kind	不适用于磁序	no
文献代理、PaperQA2 和 OpenScholar [20, 21]	yes	否，自由文本	yes	no
数据库检查代理、LitMOF [26]	测量值结构，而非磁序	而非磁磁序	仅每个条目的来源论文	no
本文工作 MagTrace	yes	yes	yes	yes

对语言模型而言，这类判断普遍很困难：它们经常依据后来已被修正的医学结论给出答案 [27]，；即使搜索代理已检索到回答问题所需的全部信息，也常会受到相互矛盾的证据误导 [28]。在磁序方面，与我们最接近的研究是 LLaMP。它会询问语言模型某种材料的磁有序情况，但只根据计算记录对磁序类型进行评分 [29]。

本文介绍了 MagTrace，这是一种大语言模型智能体；给定材料的名称后，它会决定使用哪些工具 [30]，例如文献检索或论文阅读器，并返回该材料中发现的磁序列表，同时为每个磁序提供置信度分数及支持该磁序的论文。它的工作方式类似于材料科学家手动检索文献的过程。我们的领域专用智能体会报告每个磁序及其传播矢量和引文，根据实验文献进行评估，并必须说明多个已发表磁序中哪一个是当前磁序。 **表 1** 将以往方法与我们的方法并列比较。磁序是检验此类智能体的一项严苛测试，因为论文会以许多等价方式表述同一个磁序，而已发表的磁序有时还会被修订。我们在一个包含 18 种材料的评估集上衡量 MagTrace 的表现。该评估集基于已发表论文构建，涵盖不同类型的材料及数量各异的已发表证据。

2 方法

MagTrace 模拟科学家手动从文献中核查材料磁序的过程。它接收材料的名称；语言模型（`claude-sonnet-5`）随后分轮次运行，在每一轮中，它可以使用一组工具搜索四个文献数据库中关于该材料的论文，检查论文是否已撤稿，阅读论文的免费全文，并比较磁序的不同描述。答案是当前文献所支持的磁序列表；每种磁序都会给出其有序类型、传播矢量、晶胞以及论文所述的其他细节，同时附有一个介于 0 和 100 之间的置信度分数，表示智能体对所发现的磁序当前有效且正确的置信程度，以及每个数值背后的引文（**图 1**）。完整的输出集旨在确保在获取后，只要给定晶体结构，即可为任何密度泛函理论（DFT）软件提供构建正确晶胞和磁序构型所需的全部磁性输入。

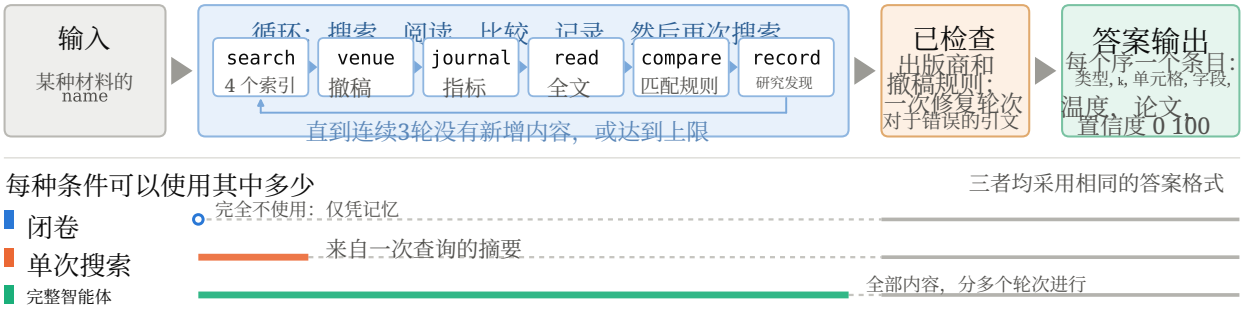

图1： 完整智能体，从左到右阅读。以材料名称作为输入，循环会进行搜索、审核、阅读、比较和记录，直到模型给出答案或代码停止循环；随后代码应用出版商规则和撤稿规则，并输出答案。下方各行显示评估（§2.3.1）的每个条件可以使用循环的多少部分。

2.1 智能体设计

智能体的主循环包含六个工具，如**表 2**所示：其中两个用于查找和阅读论文，两个用于检查每篇论文的期刊和撤稿状态，另外两个用于比较并跟踪所发现的磁序。附录 B 详细介绍了每个工具。除材料名称外，用户还可以在输入中选择智能体运行所使用的语言模型；主要结果中的所有运行均使用 `claude-sonnet-5`，附录 C列出了所有输入及示例。

每种材料的循环最多允许纳入 80 篇论文、进行 16 轮次，并读取 20 篇全文；纳入 80 篇论文后，搜索即停止，但仍可继续解读和比较。除这些上限外，代码还会在两种情况下结束一次运行：尚未记录任何序时，如果连续三次搜索都没有找到关于该材料的新论文，运行便会结束（即提前停止）；当模型记录了某种序后，如果连续三个轮次都没有发现新的序，运行便会结束（即停止规则）。如果模型的答案包含两种存在争议、且无法同时描述该材料在环境条件下状态的磁序，循环会让模型再进行一次搜索和一轮解读，然后再作答（即冲突轮次）。

生成最终输出前，代码会核查答案所引用的每篇论文的出版商和撤稿状态；如果答案引用了智能体从未检索到的论文，或引用了这些规则排除的论文，循环会将答案返回给模型一次，以便替换这些引文。附录 B 列出了具体规则及每条规则的使用频次。

每个报告的磁序都有一个置信度分数 p，范围从 0 到 100。模型根据一系列标准设定该分数：最重要的因素是独立研究团队对磁序的验证程度或一致程度（因此，单次且未经重复验证的报告不应获得高置信度分数）；次要因素包括发表时间新近度和期刊质量；第三层级的标准包括作者所属机构、资金来源、引用次数以及论文是否经过同行评审，这些因素只会略微调整分数（附录 B）。模型还可以将某种磁序标记为 存在争议，表示模型判断当前文献对该磁序存在争议。

2.2 匹配规则

论文对同一种磁序的描述往往各不相同：它们可能在不同晶胞中写出传播矢量，选择不同但对称等价的矢量，或者采用不同记法书写对称性标签。智能体需要知道两篇论文报告的是否是同一种磁序，以便统计独立确认并发现分歧；我们也需要用同样的判断标准，根据答案键评估答案。一条规则，即匹配规则，在这两个环节中作出这一判断： `compare_orders` 在循环中应用该规则，评分器则在评分时应用该规则，从而在评估中，智能体是按照实现时被要求遵循的标准进行评分的。

一般规则是：当两个磁序具有相同的有序类型，且以相同的晶胞表示时具有相同的传播矢量，它们就匹配。仅相差一个倒格矢（不会改变图案的位移）、整体符号或对称等价矢量的选取的向量，仍然

表 2： 智能体主循环中的六种工具。

工具名称	功能
`search_paper`	同时查询 Semantic Scholar、 [31], Crossref、 [32], OpenAlex [33] 和 arXiv，去除重复项，剔除可疑出版商的论文以及不属于英语的论文，并且每次调用返回 10 篇论文。
`check_venue_status`	报告论文的出版类型、出版商和撤稿状态。
`get_journal_quality`	报告期刊的引文指标。
`read_paper`	仅获取找到的论文的免费全文，并向模型传入每篇论文最多 8,000 个字符的磁学相关内容。
`compare_orders`	对关于序的两段描述应用§2.2 中的匹配规则，评判者与评分器所用的相同。
`record_candidate_orders`	让模型写下它发现的磁序，以便代码能够统计新增的磁序，用于停止规则。

匹配。在外加场下、压力下或在掺杂样品中测得的磁序，与环境条件下的磁序不同，只有在条件相同的情况下才与另一个磁序匹配。附录D 完整阐述了这条规则，包括如何处理晶胞、转变温度以及未报告传播矢量的铁磁体。

有些匹配判定过于多样，无法确定性地作出；这类判定由一个较小的模型完成，`claude-haiku-4-5`。在智能体循环和评分过程中，可以调用四个这样的模型裁判，但只有在代码发现冲突且无法自行判定时才会调用。 晶胞判定器 用于判断两个晶胞是否为同一晶胞、只是采用了不同的设置（等价的坐标轴选择）；温度判定器用于判断两个相互冲突的相变温度是否描述了同一个相；条件判定器 用于判断某个磁序所述的条件描述的是外加场、压力还是成分变化；裁决者 用于判断两个磁序是否相同，例如两个对称性标签采用了不同的记法。第五个判定器——引文判定器——仅在评估时调用，用于依据评分标准（§2.3.3）判断被引用的摘要是否支持该主张。附录 E 介绍了如何根据已知答案检验各个判定器。

2.3 评估

为了解搜索和解读文献能带来多大帮助，我们使用相同的语言模型，将 MagTrace 与两种更简单的方法进行比较，形成三个条件（§2.3.1）；测试对象为 18 种材料，其关键答案由我们根据已发表论文汇编而成（§2.3.2）。每个答案都按一个由四个部分组成的分数进行评分（§2.3.3）。

2.3.1 三个条件

这三个条件旨在检验智能体的表现：它们的区别仅在于模型能获取哪些证据，以及系统会向模型说明这些证据的哪些信息，因此条件间准确率的差异源于证据。具体如下：**闭卷**，即模型仅凭记忆作答，无法访问论文副本或智能体工具。**单次搜索**，即使用查询词“材料 磁结构”进行一次未经任何优化的搜索：将该查询发送至 OpenAlex 数据库，要求返回最多 80 篇论文，并将返回的摘要（每种材料的中位数为 67 篇）完整放入提示词中；随后模型作答一次。**完整智能体**，即 MagTrace 本身：在每一轮中，模型都可以调用 §2.1中的工具；模型给出答案或代码使其停止时，流程便结束。

其他方面在这三种条件下都相同：模型（`claude-sonnet-5`）及其设置、每个答案都必须填写的字段、关于如何权衡证据和分配置信度的说明、匹配规则和评分器（附录 F）。唯一的其他差异是，无法查阅论文的闭卷条件在给出引文前无法核查引文；此外，只有完整智能体被告知要排除潜在的掠夺性期刊和撤稿论文。每种材料在每种条件下均运行 6 次。我们将某种材料在某种条件下的一次答案称为一次运行，而所有材料和条件的第 j次运行共同构成一次重复 j。

2.3.2 评估集与答案键

评估集包含 18 种材料，其间共有 22 个磁序：每种材料各有一个当前磁序；NiI_2, 有两个有序相，因此还有第二个当前磁序；此外，3 种材料各有一个较早的、已被取代的 磁序（UO_2, $Na_2Co_2TeO_6$ 和 $CoPS_3$），后续研究修订了这些材料的磁序。我们从已发表论文中汇编了每个值，并记录其对应的论文和页码；没有任何值来自数据库条目（附录 G）。如果来源中没有说明某个字段，就将其留空且不计分，因此来源中的缺失不会降低答案的分数。

2.3.3 评分

“关键答案”描述了有关某种材料磁序的真实数据。总分为四个部分的加权和，每部分的得分范围均为 0 到 1：

- **正确性** （45%）：报告的磁序是否与关键答案中的磁序一致，并根据答案所标注的置信度分数加权。

- **细节得分** （25%）：关键答案所述的细节中，有多少也被匹配的答案正确陈述。
- **排名第一** （15%）：置信度最高的磁序是否与关键答案中的当前磁序相匹配。
- **引文完整性**（15%）：被引论文是否存在、是否符合资格（未被撤稿，也非来自可疑出版商），以及是否支持该主张。

附录H给出了这四项的公式。

正确性。每个报告的磁序至多与一个关键顺序配对。匹配的磁序得满分；仅指出序的类型（仅类型匹配）的磁序，例如只说“反铁磁”而未给出传播矢量，则得一半分。正确性将答案找到关键答案的程度（召回率）与答案所报告内容的正确程度（精确率）结合起来，并按置信度加权 F_1：每个未能匹配任何关键顺序的报告磁序都会降低精确率，且表述得越有把握，扣分越多。如果未匹配的磁序只是重述已匹配的磁序，或者描述的是关键答案未涵盖条件下的材料（例如高压或单层），且有被引摘要支持，则不扣分；我们称之为其他条件规则。

细节得分。 对于每个匹配的有序项，细节得分会依据关键答案检查四个层级：有序类型的命名有多精确（例如，如果关键答案写明“锯齿形反铁磁”，那么仅回答“反铁磁”就不够精确）、传播矢量、晶胞和对称性标签，其权重分别为 40%、32%、16% 和12%。关键答案未提及的层级会被跳过。由于细节得分仅针对匹配的有序项计算，因此材料上没有任何匹配项的条件在该材料上不会获得细节得分（附录I）。

引文完整性。 每条引文分两步核查。首先，我们检查论文是否存在以及是否符合资格。然后，我们请引文判定器判断论文摘要是支持、反驳还是未涉及该主张；根据摘要逐条判定引文，可确保对开放获取论文和付费墙论文采用同样严格的核查标准。引用一篇真实论文的 DOI，且其摘要支持该主张，可获得满分；而引用不存在、不符合资格或摘要内容与主张相矛盾的论文，则会将该有序项的引用分数设为 0，无论该有序项引用了其他什么内容（附录 J）。

所有分数均根据保存的答案计算，每位评审者都根据其保存的判定结果重新进行评估。我们报告 6 次重复实验的平均值，以及 6 次重复试验总分的标准差。§3 中的每项结果还附有 95%置信区间，该区间通过对材料进行重抽样得出，反映了结果在不同材料集上可能发生的变化（附录 K）。

3 结果

我们的结果从四个方面比较这三个条件：总分及构成总分的四个部分（§3.1）、各条件引用的论文（§3.2）、各条件针对每种材料的正确程度（§3.3），以及各条件区分当前排序与已被取代的排序、并给出与其准确率相匹配的置信度分数的能力

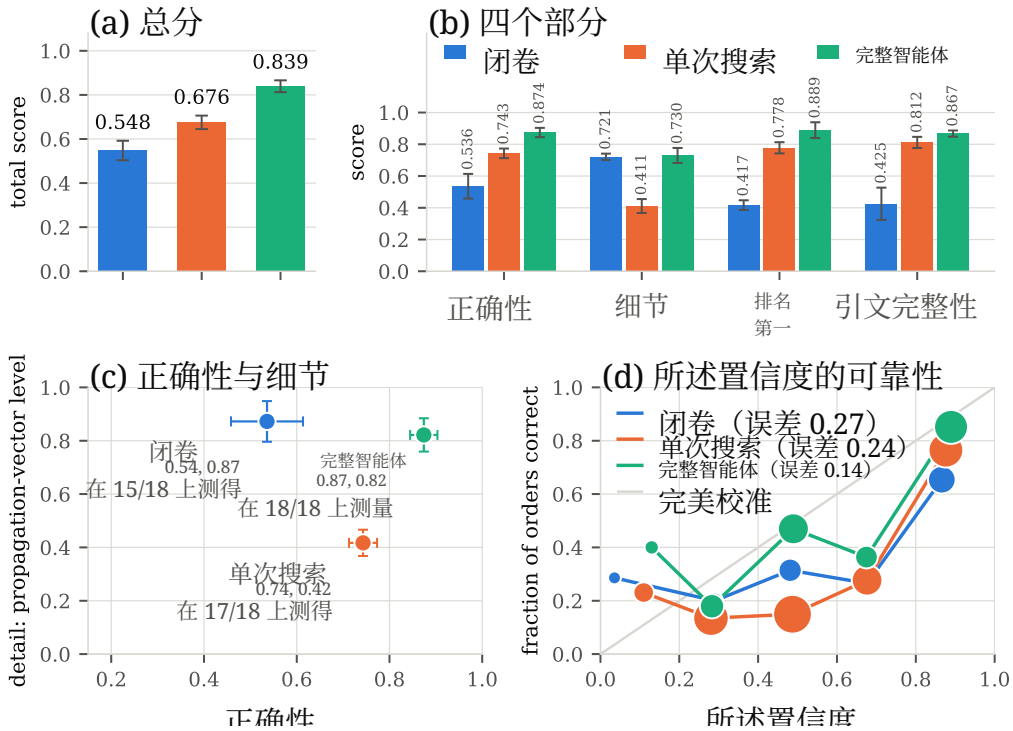

图 2： 三个条件。(a) 总分和 (b) 总分的四个部分；柱状图表示 6 次重复实验的平均值，误差线表示重复实验间的一个标准差。(c) 正确性与细节得分的传播矢量层级之间的关系（匹配的有序项中，传播矢量与关键答案一致的比例）；每个条件旁标注了可计算其细节得分的材料数量。(d) 校准：给出的置信度与该置信度下磁序正确比例之间的关系，每个20置信度区间对应一个数据点；较大的标记表示更多磁序，位于对角线下方的数据点意味着该条件过度自信。图例中的“误差”指期望校准误差（§3.4）。

（§3.4）。对于总分以及各条件之间的提升幅度，方括号中的数字表示95%置信区间，反映了结果在另一组由18种材料组成的数据集上可能有多大变化；**附录K**中的表格9列出了本节其他所有数值的置信区间。

3.1 三个条件的比较

MagTrace 在三个条件中得分最高。随着模型获得更多证据，总分不断上升：闭卷条件下为 0。548 [0.429，0。655]；单次搜索条件下为 0。676 [0.607，0。740]；完整智能体条件下为 0。839 [0.798，0。881]（**图 2a**）。因此，总分分两步以相近幅度上升：模型从凭记忆作答转为阅读单次搜索的摘要时，上升 0。128 [0.016，0。255]；从单次搜索转为完整智能体时，又上升 0。163 [0.102，0。229]。这两个区间都不包含零，因此这些提升不太可能是由我们选取的特定材料造成的。在我们测试的四个评分部分的全部 1771 种加权方案中（定义见 §2.3.3；每种组合以 5%为步长），完整智能体的得分也都最高，因此它的领先优势并不取决于我们选用的权重（附录 H）。

四个部分展示了总分的每一项提升来自何处（**图 2b**）。作为分数中占比最大的部分，正确性与总分一样分两步提升：从闭卷时的 0.536 升至单次搜索时的 0.743，再升至完整智能体时的 0.874。Top-1准确率（即置信度最高的排序是否为关键答案中的当前排序）从 0.417 升至 0.778 和 0.889；引文完整性则从 0.425 升至 0.812 和 0.867。这两部分的提升大多来自单次搜索，完整智能体带来的进一步提升较小；对于引文完整性，这一提升处于不确定性范围内（该提升的区间为 [-0.014,0.139]）。

细节得分衡量关键答案中一条匹配的磁序正确陈述了多少细节，并据此以不同顺序排列各条件：完整智能体的得分为 0.730，闭卷为 0.721，单次搜索为

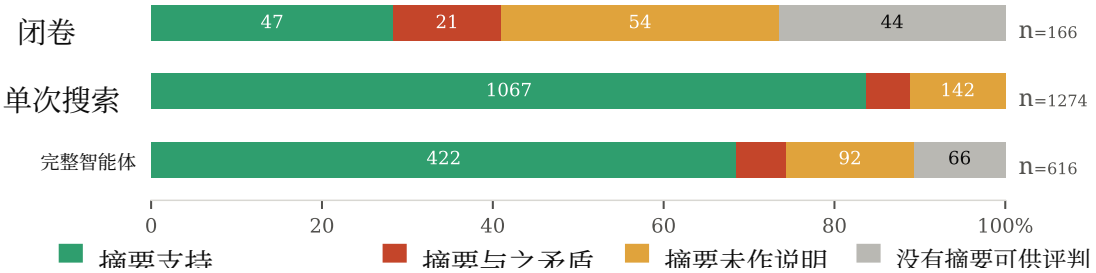

图 3: 引文判定器按条件比较每篇被引论文的摘要与引用该论文所支持的主张后，得出的结果。计数单位是引文，而非磁序。

0.411，因此完整智能体与闭卷的表现相当（两者差值的区间为 [-0.137，0。138]）。因此，尽管单次搜索的答案正确率远高于闭卷，其提供的细节却是三者中最少的。差异体现在传播矢量上（附录H，**表格7**）。单次搜索有 85% 的磁序未提及矢量，而闭卷为 32%，完整智能体为 50%：模型只阅读摘要时，能找到正确类型的磁序，却不会给出其矢量。闭卷的细节得分较高，则是另一个原因所致。只有当某种材料满足条件且至少匹配一条磁序时，才能计算其细节得分；闭卷仅在 18 种材料中的 15 种上满足这一条件（单次搜索为 17 种，完整智能体则涵盖全部 18 种）。因此，其细节得分是对模型本来就熟悉的材料取平均值（**图 2c**）。

3.2 引文

搜索论文能显著提高引文的可靠性。引文完整性用于检查被引论文是否存在、是否符合条件以及是否支持该主张；对于闭卷，其数值为 0.425，而单次搜索为 0.812，完整智能体为 0.867（**图 2b**）。闭卷的引文以两种方式失效。首先，闭卷通常仅凭标题引用论文；对于其 25% 的可检查磁序（至少有一条可查找的引文的磁序），所有引用的标题都无法对应到文献数据库中找到的论文。其次，当被引论文确实存在且有摘要时，对于闭卷的 17% 引文，摘要与该主张相矛盾；单次搜索为 5%，完整智能体为 7%（**图 3**）。由于这些论文确实存在，仅询问被引论文是否存在的检查会接受它们。 **附表 8** 在附录 J 中按磁序而非按引文给出了相同的比较结果。

图 3 按引文判定器的判定结果，细分了各条件下的引文。对于进行搜索的两个条件，大多数引用的摘要都支持该主张；在其余摘要中，未涉及该主张的摘要多于与之矛盾的摘要。闭卷条件下的引文则有所不同：支持该主张的不到一半，许多被引论文没有可供评审者阅读的摘要，而相矛盾的引文在条形图中所占比例更大：在有摘要的引文中，闭卷为 122 条中的 21 条，而单次搜索为 1274 条中的 65 条，完整智能体为 550 条中的 36 条。

3.3 各材料的正确性

正确性与细节的平均值掩盖了闭卷和单次搜索条件在不同材料之间的巨大差异。对于单种材料，闭卷的正确性范围为 0至 1，单次搜索也是如此；而完整智能体在每种材料上的正确性都高于三分之二（**图 4**）。仅统计 95%置信区间不重叠的材料时，完整智能体在 18 种材料中的 6 种上，正确性在三个条件中最高（Ba_2IrO_4, Ba_2YReO_6, CrSBr、$FePS_3$, $FePSe_3$ 和 $Na_2Co_2TeO_6$），闭卷则在 1 种材料（$CrSiTe_3$）上最高。在其余 11 种材料上，最高条件的置信区间与至少一个其他条件的置信区间重叠，因此无法判定哪个条件最高。

在**图4**中，每一行代表一种材料，每个条件对应一个点，右侧标出了三个平均值。各行按完整智能体相较于单次搜索获益最多的材料排在最上方的顺序排列。在每一行中，完整智能体的点都偏向右侧，而闭卷的点则分布在整个范围内。相较于单次搜索，Ba_2LaRuO_6, 的提升最大，其中单次搜索在每次重复中的分数均为零，而完整智能体的平均值仅在底部的 3 种材料上低于单次搜索

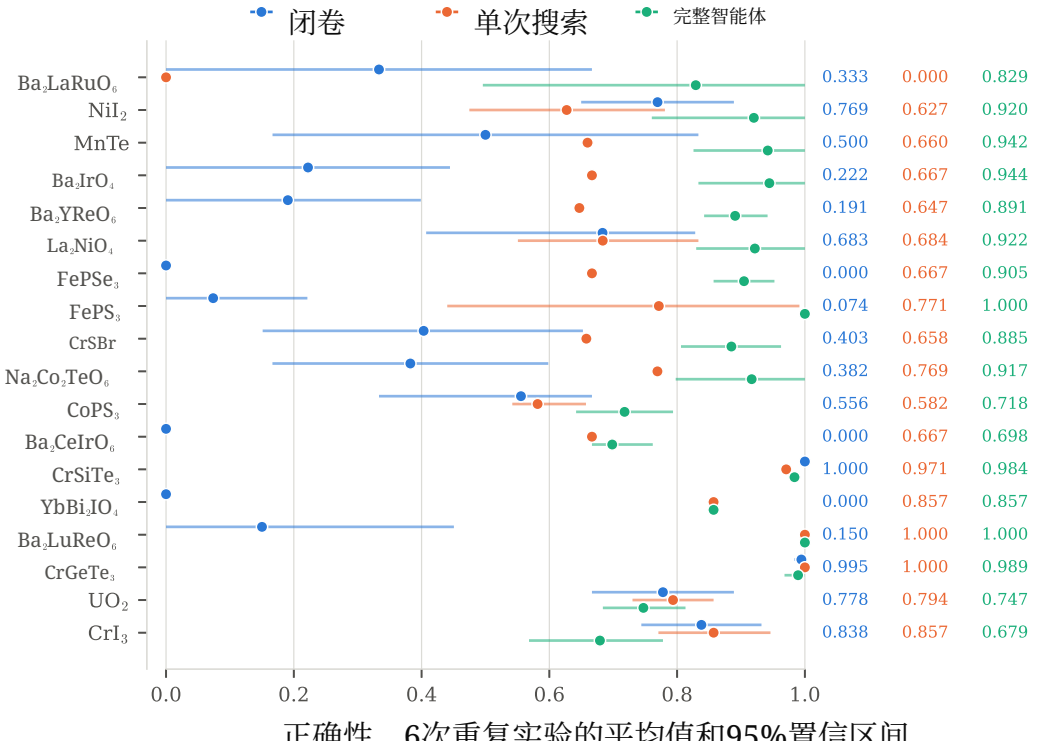

图4： 各材料的正确性，按完整智能体的正确性高于单次搜索的幅度排序，差距最大的排在最前。每个点表示6次重复实验的平均值，每条误差线表示根据这些重复实验重抽样得出的95%置信区间（附录 K）。

，$CrGeTe_3$,UO_2 和 CrI_3,的区间相互重叠。

闭卷方式也最常彻底失败，而且通常对此仍有很高的置信度。每种条件下有108个答案（18种材料，每种重复实验6次），闭卷方式有49个答案的分数恰好为零，单次搜索为7个，完整智能体则为1个。闭卷方式的零分答案中，只有6个属于弃答，即模型拒绝给出有序类型；其余43个都是有置信度的错误，即给出了错误答案且置信度大于零。例如，对于$FePS_3$ 和$FePSe_3$,，闭卷方式在每次重复实验中都给出了正确的有序类型——锯齿形反铁磁，但置信度为55至90时，它给出的传播矢量却与关键答案中的不同。有置信度的错误很危险，因为它会向读者提供错误的有序类型，却没有任何迹象表明答案需要核查。

这两个条件下的搜索也报告了远多于闭卷的磁序：单次搜索报告 372 个，完整智能体报告 229 个，而闭卷报告 158 个。额外报告的磁序中，大多数与答案键中的内容均不匹配，是答案键未收录的已发表主张：引文判定器发现，单次搜索的 255 个未匹配磁序中，有 244 个的摘要支持相应主张；完整智能体的 106 个中有 94 个，而闭卷的 95 个中只有 12 个。许多额外磁序描述的是材料在答案键所列条件以外的情形，例如高压、插层（原子或分子插入层间），或材料的单层形式（单层）。根据其他条件规则，只要引文所指的摘要支持该磁序，它就不会降低分数。附录 L 列出了未匹配磁序的具体内容，并介绍了 3 种材料，说明停止规则或我们的答案键如何影响结果。

3.4 当前磁序与已取代晶序及置信度

区分目前已确立的磁序与较早、尚未完全确定的磁序（即已取代磁序）对于保持数据库更新尤为重要；我们的评估集中有 3 种材料对此进行了测试：UO_2, $Na_2Co_2TeO_6$ 和 $CoPS_3$ 的关键答案中都同时包含当前磁序和已取代磁序。如果答案对当前磁序给出的置信度高于已取代磁序，则得 1 分；如果报告了当前磁序，但未提及已取代磁序，则得。5 分；当答案报告当前

磁序且未提及已取代磁序时，0 其他情况则得 0 分。闭卷、单次搜索和完整智能体的平均分分别为 0.33、0.69 和 0.64；由于材料数量很少，区间较宽（**表格 9**）。对于 UO_2，每种条件在每次重复中都报告了当前的三重 -k 磁序；对于 $CoPS_3$，完整智能体在每次重复中都报告了当前磁序，闭卷在 6 次中有 4 次如此，单次搜索则在 2 次中如此。$Na_2Co_2TeO_6$ 是最难的情况：通常，进行搜索的两种条件都会报告两种磁序，并将它们标记为存在争议；但在 6 个答案中，完整智能体仅有 2 个答案对当前磁序给出较高置信度，而单次搜索的 6 个答案则无一例外（附录 L）。材料仅有 3 种，因此这个分数只是一次观察，不能视为可靠比率，也不计入总分（附录H）。

由于错误的磁序陈述得越自信，正确性所受的惩罚就越大，因此置信度分数必须可靠。期望校准误差可以衡量这一点：它是答案所报告的置信度与该置信度下所报告的磁序实际正确频率之间的平均差距，因此数值越低越好。闭卷、单次搜索和完整智能体的误差分别为 0.266、0.235 和 0.135（**图 2d**）；只有闭卷与完整智能体之间的差异明显超出不确定性范围（[-0.211, -0.019]）。对于置信度最高的磁序，也呈现出相同的图案。当模型报告的置信度达到 80 或更高时，它在三种条件下的置信度大致相同：闭卷、单次搜索和完整智能体的平均置信度分数分别为 87、88 和 89。然而，这些高置信度答案的正确频率因条件而异：闭卷磁序中有 65%、单次搜索磁序中有 76%、完整智能体磁序中有 85% 与关键答案中的某个磁序匹配。由于只有 18 种材料，这些百分比存在不确定性：闭卷与完整智能体之间差异的区间为 [-2%,51%], ，其中包含零。尽管如此，单凭置信度数值，读者仍无法判断哪些答案值得信赖；而 MagTrace 的高置信度磁序与其报告的置信度最为接近。

3.5 其他智能体模型

以上所有结果均针对 `claude-sonnet-5`。由于模型是由用户选择的输入（附录 C），我们检查了完整智能体是否也能在其他 Claude 模型上正确运行。我们在 CrI_3 上分别使用九个模型（`claude-sonnet-5`、 `claude-opus-5-5`、`claude-fable-5-1`、 `claude-opus-5`、 `claude-opus-4-8`、`claude-opus-4-7`、 `claude-opus-4-6`、`claude-sonnet-4-6` 和 `claude-haiku-4-5`）各运行一次，并使用从本文一次运行中保存的工具返回结果，以确保每个模型看到相同的证据。这项检查使用六个工具，最多运行三轮，不包含附录 C 中的额外答案字段，并由 `compare_orders`仅应用固定规则。对于这九个模型，每个请求都被接受，没有答案被截断，最终答案都至少给出一个带有有序类型、置信度和引文的序；另外三个模型（`claude-sonnet-5-5`、 `claude-haiku-5-5` 和 `claude-fable-5`）无法测试，因为我们使用的模型服务不提供这些模型。这项检查表明智能体能够在这些模型上运行，但不能说明其在这些模型上的表现如何：其中，§2.3 的评估还在 `claude-opus-5-5`、`claude-fable-5-1` 和`claude-opus-5`上运行过（附录 O），而其他模型均未运行。

4 讨论

4.1 从三个条件中汲取的三点经验

我们的研究揭示了三项重要发现，适用于任何需要可信磁序的人，无论是用于自己的工作，还是用于填充或纠正数据库。

首先，我们不能相信模型的记忆。闭卷回答并非只是比另外两种条件下的回答准确率低，而是错得信心十足。在其得分为零的 49 个答案中，有 43 个以非零置信度给出了错误的磁序；而在置信度达到 80 或以上时，闭卷回答的磁序正确率仅为 65%，MagTrace 则为 85%（不过，由于涉及 18 种材料，这一差异的置信区间包含零）。闭卷回答所附的引文让情况更糟：对于 25% 附有引用标题且可核查的闭卷磁序，在文献数据库中找不到所引论文。然而，确认被引论文确实存在还不够，因为真实存在的论文也可能与其被引用来支持的主张相矛盾。被引论文存在时，其摘要有 17% 的情况下与闭卷回答的主张相矛盾，MagTrace 则为 7%。这尤其令人担忧，因为如果一项主张附有真实参考文献，读者可能更容易接受它，而不去核实该论文是否真的支持所述磁序；若没有参考文献，读者则不太容易接受。

其次，在科学数据库中进行简单搜索，可以找到磁序类型，却无法找到其细节。单次搜索的正确率远高于闭卷回答，但它提供的细节却是三种方式中最少的：其给出的磁序有 85% 未提供传播矢量。只说“反铁磁”而不提供矢量，答案就无法与某个已发表的有序态对应，无法据此开始计算，也无法与论文报告的矢量进行核对。如果读者只看到正确性得分，就会认为单次搜索显然优于凭记忆作答；而如果只看到细节得分，则会得出相反结论。因此，我们在报告总分的同时，也列出得分的四个部分（附录 H）。

第三，每类证据都会改善答案的不同方面。正确性分两步提升：第一步是向模型提供单次搜索的摘要，第二步是模型使用完整智能体进行重复搜索和解读。引文完整性的提升主要来自单次搜索，使其从 0.425 升至 0.812；完整智能体进一步将其提升至 0.867，但这一增幅处于不确定性范围内。细节方面的提升则来自完整智能体：它提及传播矢量的频率远高于单次搜索，并使细节得分恢复到闭卷水平。因此，为这项任务构建的系统既需要检索来提供读者可核查的引文，也需要完整的搜索与解读循环，才能给出既正确又具体的答案。

4.2 MagTrace 的潜在用例

MagTrace 已可用于以下两种近期应用，其设计也为未来在三个方向上的拓展留有余地。扩大答案键，纳入更多排序后来经过修订的材料（我们的材料中只有 3 种），并为我们的来源未说明晶胞的 4 种材料补充注明晶胞的来源，就能更准确地衡量智能体区分当前排序与已被取代排序的能力。完整智能体的置信度分数比闭卷方式校准得更好（期望校准误差为 0.135，而闭卷方式为 0.266）；若要按智能体对数据库条目与文献不一致的置信度进行排名，这一点至关重要。这样的排名可帮助数据维护者找出最需要核查的条目，而对其进行测试是顺理成章的下一步。最后，这一分数体系可扩展至用户自身问题所关注的内容，例如相变温度或各相的温度范围，而智能体已经能够报告这些信息：添加新标准只需声明新的答案字段、用于比较该字段的规则，以及该字段在分数中的权重（附录 H）。

交叉核查数据库 MAGNDATA 可知，其中收录了 18 种材料中的 8 种，共 13 条记录；我们将每条记录所主张的磁序与答案键以及完整智能体给出的答案进行了比较（附录 A）。对于其中 5 种材料，这些记录在传播矢量方面与答案键当前采用的磁序一致，而完整智能体在每次重复中最确信的磁序也都与之相符。其余 3 种材料有多条数据库记录，且记录之间并不完全一致，因此仅凭数据库，读者无法判断哪种磁序是当今普遍认可的材料磁序。在 $Na_2Co_2TeO_6$, 2的 4 条记录中，记录的都是答案键来源所述、已被取代的单$\mathbf{k}$ 锯齿型磁序；而 2 条 MnTe 记录与 2 条 La_2NiO_4 记录一样，在传播矢量方面与答案键一致，但给出的磁空间群各不相同。因此，数据库可能会并列收录同一种材料的两种磁序，却不注明哪一种已被取代；而 MagTrace 可以指出哪些论文对此作出了定论。

发现开放问题及后续步骤 一个智能体通过解读关于某种材料的文献来确定其磁序时，也能发现这些文献内部存在的分歧；两项审慎测量结果之间的分歧是一个研究问题，而不是记录中的错误。MagTrace 已经发现了这类情况。对于 $Na_2Co_2TeO_6$，它并列报告单$\mathbf{k}$ 和三重$\mathbf{k}$ 描述，并将其标记为有争议（附录 M）；对于 Ba_2YReO_6，它报告的转变温度比我们的关键答案中记录的更新；对于大多数材料，它报告了我们的关键答案未记录的已发表磁序，包括高压下、薄样品中或通过计算得出的结果（附录 L）。对于每一种磁序，每个答案都会记录论文之间存在分歧的字段，并注明各项解读及其出处论文；汇集不同材料的这些记录，就能形成一份开放问题清单。

4.3 局限性

我们的评估　　有六项主要局限，列举如下；附录M对其中三项作了更详细的说明　　.

判断某个序是当前排序还是已取代晶序，取决于 3 种材料。只有 UO_2,$Na_2Co_2TeO_6$ 和 $CoPS_3$ 的关键答案中同时列出了当前排序和已取代晶序，因此，我们对这项能力的测量最不充分，尽管它对于保持数据库内容最新至关重要。

答案键并不完整。 有 4 种材料没有任何来源给出晶胞，且只有 2 个磁序带有磁空间群，因此会跳过这些检查。对于大多数材料，答案键只收录一个磁序，因此当答案报告了答案键未列出的已发表相时，该磁序与答案键中的任何内容都不匹配，并会降低答案的正确性得分（陈述得越肯定，降低得越多），除非适用其他条件规则。在 Ba_2YReO_6 上，答案键还与一项较新的研究在转变温度方面存在分歧（附录 L），这会使完整智能体在大多数重复实验的排名第一字段中的分数降低。

评分器及其评判者。 一些评分规则是在阅读答案后制定的；附录 N列出了每条规则及其影响，而条件则按相同顺序排列，且不包含其中任何一条规则。评判者仅针对我们自己运行中的冲突进行了检查，因此尚不清楚它们在其他材料上的表现如何；在人工标注的 31 个困难案例中，引文判定器有 8 个判断错误，最常见的错误是：当摘要仅以更概括的术语描述某种磁性模式时，却接受了一个具体的磁性模式（附录 E）。

引文核查。忠实度是依据摘要来评判的，因此，只有全文才能揭穿的主张也可能过关；而在书目数据库中没有摘要的论文则完全不会接受评判。若引文仅凭一个与数据库中某篇论文相似的标题给出，却无法明确匹配该论文，就不会纳入核查：闭卷的引文中有 42 条属于这种情况，单次搜索则有 0 条（附录J）。

提前停止。连续 3 次搜索都未能发现有关该材料的新内容后便结束搜索，对于文献较少的材料可以节省精力；但提前停止也可能在关键论文出现之前就结束搜索，尤其是当论文标题未提及该材料时，例如 Ba_2CeIrO_6 （附录L）。

比较所揭示的差异。主要结果中的每种条件都在一个模型上运行， `claude-sonnet-5`；在三个较新的模型上，按照相同条件进行相同实验，结果显示完整智能体在每个模型上的平均分数都是最高的；但在其中两个模型上，它相对闭卷的领先幅度处于不确定性范围内，而且闭卷的平均分数高于单次搜索（附录O）。完整智能体与单次搜索还同时存在三方面差异（搜索四个数据库、进行多次而非单次搜索，以及阅读全文），因此我们无法确定其增益有多少分别来自这几方面。重复运行之间低于约0.05 的分数差异应视为噪声（附录K）。

5 结论

我们依据已发表论文中的答案键，对 MagTrace 在 18 份测试材料上的表现进行了评分。MagTrace 会反复搜索四个文献数据库，并阅读开放获取全文，得分为 0.839；相比之下，仅搜索摘要的简单搜索得分为 0.676，而同一模型凭记忆作答得分为 0.548。在我们尝试的评分四个部分的所有权重设置下，它都排名第一（附录 H）。单次搜索带来了引文完整性方面的大部分提升，完整智能体弥补了单次搜索所损失的细节，两者也都提高了正确性。凭记忆作答的模型经常引用并不存在的论文标题；即便论文确实存在，其摘要与主张相矛盾的情况也比两种会进行搜索的条件更常见。

数据可用性

答案键（其中标明了每个数值所依据的论文和页码）、每次运行保存的答案、每位评审者保存的判定结果以及引文查询，均包含在下方的代码仓库中，因此本文中的每个数字都可以离线重新计算。答案键所依据的论文均受版权保护，因此仅作引文引用，不予再分发。

代码可用性

智能体、评分器以及生成每个表格和图表的脚本，都位于一个将在论文发表时公开的 GitHub 代码仓库中，同时还会提供一个标记，指向生成本文所报告结果的修订版本。

致谢

斯坦福大学和SLAC国家加速器实验室的工作得到美国能源部基础能源科学办公室材料科学与工程部的支持，项目编号为DE-SCL0000172。本研究使用了国家能源研究科学计算中心的资源。该中心是美国能源部科学办公室用户设施，由美国能源部科学办公室根据合同编号DE-AC02-05CH11231提供支持，本研究使用的NERSC项目编号为BES-ERCAP0034968。M. Sun在暑期实习期间获得了SLAC国家加速器实验室直线加速器相干光源（LCLS）暑期实习计划的支持。SLAC国家加速器实验室由美国能源部科学办公室基础能源科学办公室根据合同编号DE-AC0276SF00515提供支持。本文研究的系统基于Anthropic的 `claude-sonnet-5` 和 `claude-haiku-4-5` 模型构建，具体说明见方法。除此之外，在作者指导下，还使用人工智能工具协助编写代码并修改本手稿的语言。作者对本工作的内容承担全部责任。

作者贡献

M.S. 设计并构建了智能体和评估，编制了答案键，开展了实验，分析了结果并撰写了手稿。C.P. 构思了该项目，提供了作为答案键依据的来源论文和传播矢量，审阅了答案键，并就研究工作的框架提出建议。两位作者均审阅了手稿。

利益冲突

作者声明不存在利益冲突。

参考文献

[1] Juan Rodríguez-Carvajal. 通过中子粉末衍射测定磁结构的最新进展。《物理学*B*：凝聚态》，192(1-2):55–69，1993年。doi: 10.1016/0921-4526(93)90108-i。[2] Libor Šmejkal、Jairo Sinova 和 Tomas Jungwirth。超越传统铁磁性与反铁磁性：一种具有非相对论自旋和晶体旋转对称性的相。*Physical Review X*，12(3):031042，2022年。doi: 10.1103/physrevx.12.031042。[3] B. Andrei Bernevig、Claudia Felser 和 Haim Beidenkopf。磁性拓扑材料的进展与前景。《自然》，603(7899):41–51，2022年。doi: 10.1038/s41586-021-04105-x。[4] Yuanfeng Xu、Luis Elcoro、Zhi-Da Song、Benjamin J. Wieder、M. G. Vergniory、Nicolas Regnault、Yulin Chen、Claudia Felser 和 B. Andrei Bernevig。磁性拓扑材料的高通量计算。《自然》，586(7831):702–707，2020年。doi: 10.1038/s41586-020-2837-0。[5] J.M. Perez-Mato、S.V. Gallego、E.S. Tasci、L. Elcoro、G. de la Flor 和 M.I. Aroyo。用于磁晶体学的基于对称性的计算工具。《材料研究年度评论》，45(1):217–248，2015年。doi: 10.1146/annurev-matsci-070214-021008。[6] Helena A. Merker、Harry Heiberger、Linh Nguyen、Tongtong Liu、Zhantao Chen、Nina Andrejevic、Nathan C. Drucker、Ryotaro Okabe、Song Eun Kim、Yao Wang、Tess Smidt 和 Mingda Li。机器

learni磁性 tism classifiers from a tomic coor dinates. *iSci*$_{ence}$, 25(10) :105192 d2022. oi: 10.1016/j . isci.2022.105192.[7] Samuel V. Gallego、J. Manuel Perez-Mato、Luis Elcoro、Emre S. Tasci、Robert M. Hanson、Koichi Momma、Mois I. Aroyo 和 Gotzon Madariaga。MAGNDATA：迈向磁结构数据库。I. 公度情形。《应用晶体学杂志》，49(5):1750–1776，2016年。doi: 10.1107/s1600576716012863。[8] Samuel V. Gallego、J. Manuel Perez-Mato、Luis Elcoro、Emre S. Tasci、Robert M. Hanson、Mois I. Aroyo 和 Gotzon Madariaga。MAGNDATA：迈向磁结构数据库。II. 非公度情形。《应用晶体学杂志》，49(6):1941–1956，2016年。doi: 10.1107/s1600576716015491。[9] Suman Itani、Yibo Zhang 和 Jiadong Zang。面向磁性材料的东北材料数据库。Nature Communications，16(1):9415，2025年。doi: 10.1038/s41467-025-64458-z。[10] G. H. Lander 和 R. Caciuffo。理解 UO_2 有序态动力学所耗费的五十年。Journal of Physics: Condensed Matter，32(37):374001，2020年。doi: 10.1088/1361-648x/ab1dc5。[11] Matthew Kristofer Horton、Joseph Harold Montoya、Miao Liu 和 Kristin Aslaug Persson。使用密度泛函理论高通量预测无机材料的基态共线磁序。npj Computational Materials，5(1):64，2019年。doi: 10.1038/s41524-019-0199-7。[12] Vinay I. Hegde、Christopher K. H. Borg、Zachary del Rosario、Yoolhee Kim、Maxwell Hutchinson、Erin Antono、Julia Ling、Paul Saxe、James E. Saal 和 Bryce Meredig。量化高通量密度泛函理论中的不确定性：AFLOW、Materials Project 和 OQMD 的比较。《物理评论：材料》，7(5):053805，2023年。doi: 10.1103/physrevmaterials.7.053805。[13] Matthew C. Swain 和 Jacqueline M. Cole。ChemDataExtractor：从科学文献中自动提取化学信息的工具包。《化学信息与建模杂志》，56(10):1894–1904，2016年。doi: 10.1021/acs.jcim.6b00207。[14] Maciej P. Polak 和 Dane Morgan。使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。Nature Communications，15(1):1569，2024年。doi: 10.1038/s41467-024-45914-8。[15] Callum J. Court 和 Jacqueline M. Cole。通过半监督关系抽取自动生成居里温度和奈尔温度材料数据库。《科学数据》，5(1):180111，2018年。doi: 10.1038/sdata.2018.111。[16] Yibo Zhang、Suman Itani、Kamal Khanal、Emmanuel Okyere、Gavin Smith、Koichiro Takahashi 和 Jiadong Zang。GPTArticleExtractor：用于构建磁性材料数据库的自动化工作流。《磁学与磁性材料杂志》，597:172001，2024年。doi: 10.1016/j.jmmm.2024.172001。[17] Ahmed E. Fahmy。从大规模材料数据库中学习磁序分类。《物理评论：材料》，10(7):074414，2026年。doi: 10.1103/qftr-2tfm。[18] William H. Walters 和 Esther Isabelle Wilder。ChatGPT 生成的书目引文中的编造与错误。《科学报告》，13(1):14045，2023年。doi: 10.1038/s41598-023-41032-5。[19] Jakub Lála、Odhran O'Donoghue、Aleksandar Shtedritski、Sam Cox、Samuel G. Rodriques 和 Andrew D. White。PaperQA：用于科学研究的检索增强生成智能体。arXiv:2312.07559，2023年。URL https://arxiv.org/abs/2312.07559。[20] Michael D. Skarlinski、Sam Cox、Jon M. Laurent、James D. Braza、Michaela Hinks、Michael J. Hammerling、Manvitha Ponnapati、Samuel G. Rodriques 和 Andrew D. White。语言智能体实现超越人类水平的科学知识综合。arXiv:2409.13740，2024年。URL https://arxiv.org/abs/2409.13740。[21] Akari Asai、Jacqueline He、Rulin Shao、Weijia Shi、Amanpreet Singh、Joseph Chee Chang、Kyle Lo、Luca Soldaini 等。使用检索增强语言模型综合科学文献。《自然》，650(8103):857–863，2026年。doi: 10.1038/s41586-025-10072-4。

[22] Andres M. Bran、Sam Cox、Oliver Schilter、Carlo Baldassari、Andrew D. White 和 Philippe Schwaller。使用化学工具增强大语言模型。Nature Machine Intelligence，6(5):525–535，2024年。doi: 10.1038/s42256-024-00832-8。[23] Daniil A. Boiko、Robert MacKnight、Ben Kline 和 Gabe Gomes。使用大语言模型开展自主化学研究。《自然》，624(7992):570–578，2023年。doi: 10.1038/s41586-023-06792-0。[24] Akshat Chaudhari、Janghoon Ock 和 Amir Barati Farimani。面向多任务计算材料科学的模块化大语言模型智能体。Communications Materials，7(1):131，2026年。doi: 10.1038/ s43246-025-00994-x。[25] Mehrad Ansari 和 Seyed Mohamad Moosavi。从科学文献中基于智能体学习材料数据集。Digital Discovery，3(12):2607–2617，2024年。doi: 10.1039/d4dd00252k。[26] Honghui Kim、Dohoon Kim 和 Jihan Kim。LitMOF：用于经文献验证的金属有机框架数据库校正与扩展的大语言模型多智能体。arXiv:2512.01693，2025年。URL https://arxiv.org/abs/2512.01693。[27] Juraj Vladika、Mahdi Dhaini 和 Florian Matthes。事实迅速淡忘：评估大语言模型对过时医学知识的记忆。在《计算语言学协会会议论文集：EMNLP 2025》发表，2025年。URL https://arxiv.org/abs/2509.04304。[28] Tianyi Hu、Bhavyajeet Singh、Niket Tandon 和 Akhil Arora。搜索增强型大语言模型在快速演变的知识推理中于何处以及为何失败。arXiv:2610.25323，2026年。URL https://arxiv.org/abs/2610.25323。[29] Yuan Chiang、Elvis Hsieh、Chia-Hong Chou 和 Janosh Riebesell。LLaMP：让大语言模型具备高保真材料知识检索能力。在《2025年自然语言处理实证方法会议论文集》中，25189–25221页。计算语言学协会，2025年。doi: 10.18653/v1/2025.emnlp-main.1280。[30] Shunyu Yao、Jeffrey Zhao、Dian Yu、Nan Du、Izhak Shafran、Karthik Narasimhan 和 Yuan Cao。ReAct：协同语言模型中的推理与行动。在第十一届国际学习表征会议（ICLR）上，2023年。URL https://openreview.net/forum?id=WE_vluYUL-X。[31] Rodney Kinney、Chloe Anastasiades、Russell Authur、Iz Beltagy、Jonathan Bragg 等。《Semantic Scholar 开放数据平台》。arXiv:2301.10140，2023年。URL https://arxiv.org/abs/2301.10140。[32] Ginny Hendricks、Dominika Tkaczyk、Jennifer Lin 和 Patricia Feeney。Crossref：社区所有的学术元数据的可持续来源。Quantitative Science Studies，1(1):414–427，2020年。doi: 10.1162/qss_a_00022。[33] Jason Priem、Heather Piwowar 和 Richard Orr。OpenAlex：学术作品、作者、发表平台、机构和概念的完全开放索引。arXiv:2205.01833，2022年。URL https://arxiv.org/abs/2205.01833。[34] Nobuhiko Kunitomi、Yoshikazu Hamaguchi 和 Shuichiro Anzai。碲化锰的中子衍射研究。Journal de Physique，25(5):568–574，1964年。doi: 10.1051/jphys:01964002505056800。[35] Zhaoyu Liu、Sijie Xu、Jonathan M. DeStefano、Elliott Rosenberg、Tingjun Zhang、Jinyulin Li、Matthew B. Stone、Feng Ye、Wei Tian、Sarah Edwards、Rong Cong、Siyu Pan、Ching-Wu Chu、Liangzi Deng、Emilia Morosan、Rafael M. Fernandes、Jiun-Haw Chu 和 Pengcheng Dai。六方 MnTe 中应变可调反常霍尔效应。《物理评论 X》，16(3):031033，2026年。doi: 10.1103/589s-s1yy。[36] A. A. Aczel、J. P. Clancy、Q. Chen、H. D. Zhou、D. Reig-i Plessis、G. J. MacDougall、J. P. C. Ruff、M. H. Upton、Z. Islam、T. J. Williams、S. Calder 和 J.-Q. Yan。重新审视 Ir^{4+}双钙钛矿铱酸盐的 Kitaev 材料候选资格。《物理评论 B》，99(13):134417，2019年。doi: 10.1103/physrevb.99.134417。

[37] A. Revelli、C. C. Loo、D. Kiese、P. Becker、T. Fröhlich、T. Lorenz、M. Moretti Sala、G. Monaco、F. L. Buessen、J. Attig、M. Hermanns、S. V. Streltsov、D. I. Khomskii、J. van den Brink、M. Braden、P. H. M. van Loosdrecht、S. Trebst、A. Paramekanti 和 M. Grüninger。Ba $j = 1/2$ $CeIrO_2$中的自旋轨道纠缠$_6$磁矩：一种受挫面心立方量子磁体。Physical Review B，100(8):085139，2019年。doi：10.1103/physrevb.100.085139。[38] S. Boseggia、R. Springell、H. C. Walker、H. M. Rønnow、Ch. Rüegg、H. Okabe、M. Isobe、R. S. Perry、S. P. Collins 和 D. F. McMorrow。单层铱酸盐自旋轨道莫特绝缘体中基面反铁磁序及 $J_{\text{eff}} = 1/2$态的稳健性。Physical Review Letters，110(11):117207，2013年。doi：10.1103/physrevlett.110.117207。[39] J. P. Clancy、H. Gretarsson、A. Lupascu、J. A. Sears、Z. Nie、M. H. Upton、Jungho Kim、Z. Islam、M. Uchida、D. G. Schlom、K. M. Shen 和 Young-June Kim。方格晶格铱酸盐 Ba_2IrO_4中的磁激发。Physical Review B，107(5):054423，2023年。doi：10.1103/physrevb.107.054423。[40] P.D. Battle、J.B. Goodenough 和 R. Price。Ba_2LaRuO_6 和 Ca_2LaRuO_6的晶体结构与磁性。Journal of Solid State Chemistry，46(2):234–244，1983年。doi：10.1016/ 0022-4596(83)90147-0。[41] P.D. Battle 和 C.W. Jones。Sr_2LuRuO_6，、Ba_2YRuO_6，和 Ba_2LuRuO_6的晶体结构与磁结构。Journal of Solid State Chemistry，78(1):108–116，1989年。doi：10.1016/0022-4596(89)90133-3。[42] Jie Xiong、Jiaqiang Yan、Adam A. Aczel 和 Patrick M. Woodward。含 $5d^2$ 离子的双钙钛矿 Ba_2LuReO_6中的 I 型反铁磁序：探索结构畸变的作用。Journal of Solid State Chemistry，258:762–767，2018年。doi：10.1016/j.jssc.2017.11.040。[43] Gøran J. Nilsen、Corey M. Thompson、Casey Marjerisson、Danis I. Badrtdinov、Alexander A. Tsirlin 和 John E. Greedan。 $5d^2$ 铼双钙钛矿 Ba_2YReO_6中的磁序与多极矩。Physical Review B，103(10):104430，2021年。doi：10.1103/physrevb.103.104430。[44] A R Wildes、V Simonet、E Ressouche、R Ballou 和 G J McIntyre。准二维反铁磁体 $CoPS_3$的磁性与结构。Journal of Physics: Condensed Matter，29(45):455801，2017年。doi：10.1088/1361-648x/aa8a43。[45] Fengfeng Zhu、Lichuan Zhang、Xiao Wang、Flaviano José dos Santos、Junda Song、Thomas Mueller、Karin Schmalzl、Wolfgang F. Schmidt、Alexandre Ivanov、Jitae T. Park、Jianhui Xu、Jie Ma、Samir Lounis、Stefan Blügel、Yuriy Mokrousov、Yixi Su 和 Thomas Brückel。二维范德华铁磁体 $CrSiTe_3$ 和 $CrGeTe_3$中的拓扑磁振子绝缘体：迈向本征带隙可调性。Science Advances，7(37):eabi7532，2021年。doi：10.1126/sciadv.abi7532。[46] Michael A. McGuire、Hemant Dixit、Valentino R. Cooper 和 Brian C. Sales。层状铁磁绝缘体 CrI_3中晶体结构与磁性的耦合。Chemistry of Materials，27(2): 612–620，2015年。doi：10.1021/cm504242t。[47] Liqin Ke 和 Mikhail I. Katsnelson。电子关联效应对二维范德华材料中交换相互作用和自旋激发的影响。npj Computational Materials，7(1):4，2021年。doi：10.1038/ s41524-020-00469-2。[48] Allen Scheie、Michael Ziebel、Daniel G. Chica、Youn June Bae、Xiaoping Wang、Alexander I. Kolesnikov、Xiaoyang Zhu 和 Xavier Roy。CrSBr中的自旋波与磁交换哈密顿量。Advanced Science，9(25):2202467，2022年。doi：10.1002/advs.202202467。[49] Matthew J. Coak、David M. Jarvis、Hayrullo Hamidov、Andrew R. Wildes、Joseph A. M. Paddison、Cheng Liu、Charles R. S. Haines、Ngoc T. Dang、Sergey E. Kichanov、Boris N. Savenko、Sungmin Lee、Marie Kratochvílová、Stefan Klotz、Thomas C. Hansen、Denis P. Kozlenko、Je-Geun Park 和 Siddharth S. Saxena。压力调控的范德华反铁磁体 $FePS_3$中的涌现磁相。Physical Review X，11(1):011024，2021年。doi：10.1103/physrevx.11.011024。

[50] Amber McCreary、Jeffrey R. Simpson、Thuc T. Mai、Robert D. McMichael、Jason E. Douglas、Nicholas Butch、Cindi Dennis、Rolando Valdés Aguilar 和 Angela R. Hight Walker。通过磁拉曼光谱识别反铁磁 $FePS_3$ 中的准二维磁振子。Physical Review B，101(6):064416，2020年。doi：10.1103/physrevb.101.064416。[51] Mohammad Amirabbasi 和 Peter Kratzer。单层 $FePS_3$中的轨道有序与磁有序：一项 DFT+U 研究。Physical Review B，107(2):024401，2023年。doi：10.1103/physrevb.107.024401。[52] Lebing Chen、Xintong Li、Zhenhai Hu、Ze Hu、Li Yue、Ronny Sutarto、Feizhou He、Kazuki Iida、Kazuya Kamazawa、Weiqiang Yu、Xi Lin 和 Yuan Li。蜂窝伊辛反铁磁体 $FePSe_3$中自旋激发的热演化。npj Quantum Materials，9(1):40，2024年。doi：10.1038/s41535-024-00651-5。[53] Rabindra Basnet、Taksh Patel、Jian Wang、Dinesh Upreti、Santosh Karki Chhetri、Gokul Acharya、Md Rafique Un Nabi、Josh Sakon 和 Jin Hu。为潜在二维磁体理解并调控层状伊辛型反铁磁体 $FePSe_3$ 中的磁性。Advanced Electronic Materials，10(7): 2300738，2024年。doi：10.1002/aelm.202300738。[54] A. N. Petsch、N. S. Headings、D. Prabhakaran、A. I. Kolesnikov、C. D. Frost、A. T. Boothroyd、R. Coldea 和 S. M. Hayden。自旋-1正方晶格反铁磁体 La_2NiO_4中的高能自旋波。Physical Review Research，5(3):033113，2023年。doi：10.1103/physrevresearch.5.033113。[55] I. I. Mazin。MnTe中的交错磁性：起源、预测表现与去孪晶途径。Physical Review B，107(10):L100418，2023年。doi：10.1103/physrevb.107.l100418。[56] O. J. Amin、A. Dal Din、E. Golias、Y. Niu、A. Zakharov、S. C. Fromage、C. J. B. Fields、S. L. Heywood、R. B. Cousins、F. Maccherozzi、J. Krempaský、J. H. Dil、D. Kriegner、B. Kiraly、R. P. Campion、A. W. Rushforth、K. W. Edmonds、S. S. Dhesi、L. Šmejkal、T. Jungwirth 和 P. Wadley。MnTe中交错磁性的纳米尺度成像与调控。《自然》，636(8042):348–353，2024年。doi：10.1038/s41586-024-08234-x。[57] Wenjie Chen、Xintong Li、Zhenhai Hu、Ze Hu、Li Yue、Ronny Sutarto、Feizhou He、Kazuki Iida、Kazuya Kamazawa、Weiqiang Yu、Xi Lin 和 Yuan Li。低温下 $Na_2Co_2TeO_6$ 的自旋轨道相行为。Physical Review B，103(18):L180404，2021年。doi：10.1103/physrevb.103.l180404。[58] Anjana M. Samarakoon、Qiang Chen、Haidong Zhou 和 V. Ovidiu Garlea。蜂窝晶格反铁磁体 $Na_2M_2TeO_6$, M =Co 和 Ni 的静态与动态磁性。Physical Review B，104 (18):184415，2021年。doi：10.1103/physrevb.104.184415。[59] S.R. Kuindersma、J.P. Sanchez 和 C. Haas。NiI_2 和 CoI_2的磁性与结构研究。Physica B+C，111(2-3):231–248，1981年。doi：10.1016/0378-4363(81)90100-5。[60] Qiye Liu、Wenjie Su、Yue Gu、Xi Zhang、Xiuquan Xia、Le Wang、Ke Xiao、Naipeng Zhang、Xiaodong Cui、Mingyuan Huang、Chengrong Wei、Xiaolong Zou、Bin Xi、Jia-Wei Mei 和 Jun-Feng Dai。NiI_2中从螺旋磁有序到反铁磁态的意外压力诱导磁性转变。Nature Communications，16(1):4221，2025年。doi：10.1038/s41467-025-59561-0。[61] Joseph A. M. Paddison、Lionel Desgranges、Gianguido Baldinozzi、Gerard H. Lander 和 Henry E. Fischer。自旋轨道耦合磁体 UO_2中的 Kitaev相互作用。arXiv:2512.19674，2025年。URL https://arxiv.org/abs/2512.19674。[62] Pyeongjae Park、Qianli Ma、G. Sala、S. Calder、Douglas L. Abernathy、Matthew B. Stone、Andrew F. May 和 Andrew D. Christianson。具有 $J_{\text{eff}} = 1/2$的正方晶格反铁磁体 $YbBi_2ClO_4$ 和 $YbBi_2IO_4$ 的晶场劈裂与磁基态。Physical Review B，111(10):104430，2025年。doi：10.1103/physrevb.111.104430。[63] Otkur Omar、Yang Zhang、Qiang Zhang、Wei Tian、Elbio Dagotto、Gang Chen、Taka-hisa Arima、Matthew B. Stone、Andrew D. Christianson、Daigorou Hirai 和 Shang Gao。 $5d^2$ Re基双钙钛矿 Ba_2YReO_6 和 Ba_2ScReO_6中的偶极与四极关联。Physical Review B，112 (7):075103，2025年。doi：10.1103/gwph-w9zm。

表3： 我们的材料对应的 magndata 条目，与答案键和完整智能体并列展示。“智能体”总结了完整智能体在 6 次重复实验中最有把握的序。条目编号旁的引文是该条目引用的论文；主张则是条目自身的主张。

材料，条目	条目的主张	与关键答案和智能体对照
UO_2, 3.2	三重-**k**，⟨100⟩ 星形，$Pn\bar{3}m'$	与当前排序一致；智能体的最高排序每次重复时都匹配其三重**k** 序，但名称为 $Pa\bar{3}$ 而非 Pn $\bar{3}m'$ 在 6 次中的 4 次
$FePS_3$, 1.183	$\mathbf{k}=(0,\ 1,\ \perp_2)$ 位于 $C2/m$	一致；智能体每次重复时都一致
$FePSe_3$, 1.210	$\mathbf{k}=(\frac{1}{2},0,\frac{1}{2})$, 在 $R\bar{3}$ 中	一致；智能体每次重复时都一致
CrSBr，1.802	$\mathbf{k}=(0,0,\perp_2)$, 在 $Pmmn$ 中	一致；智能体在每次重复实验中都一致
$CoPS_3$, 1.264	$\mathbf{k}=(0,1,0)$ 在 $C2/m$, P_C2_1/m	与当前排序一致，且来自与……相同的论文 关键答案；智能体在 6 次重复实验中的 5 次给出名称 P_C2_1/m
$Na_2Co_2TeO_6$, 1.184 和 1.645	单一-**k** $(\frac{1}{2},0,0)$, $P_C2_12_12_1$	关键答案已被取代的锯齿型磁序
$Na_2Co_2TeO_6$, 3.29	三重-**k**，$P6_32'2'$	与当前排序一致
$Na_2Co_2TeO_6$, 1.771	母相 $C2/m$, $\mathbf{k}=(\frac{1}{2},\frac{-1}{2},0)$	另一种（单斜）多晶型；不可比较
MnTe, 0.800 [34]	$\mathbf{k}=0$, $Cmcm$	在 **k** 上一致；关键答案未给出磁空间群
MnTe，0.1155 [35]	$\mathbf{k}=0$，$Cm'c'm$	在**k**上达成一致；智能体名称为 $Cm'c'm$ ，在 6 次重复实验中有 3 次
La_2NiO_4，0.45 以及 1.42	$Pc'c'n$，$\mathbf{k}=0$；P_Cmna，$\mathbf{k}=(1,\ 0,\ 0)$ 在 $Bmeb$	两者对**k**的看法一致：关键答案的（$\perp_2$，$\perp_2$，0）在 $Bmeb$ 中为 (1，0，0) 以及一个晶格向量，因此在 $P4_2/ncm$ 的单元格中，$\mathbf{k}=0$ 0.45；关键答案给出的是不可约表示，而不是磁空间群；智能体给出关键答案的面内 $(\perp_2,\perp_2)$ 在每个重复周期中

A 数据库交叉核查

表格 3 将每个 MAGNDATA 条目的主张与我们的答案键以及完整智能体的答案并列。仅当条目和答案键使用相同的母设定时，才会比较传播矢量；若答案键未提供可供比较的信息，或条目描述的是不同的多晶型，则将该条目标记为不可比较。条目中的数值是其自身的主张，取自条目本身，不会用于答案键或评分。

整个数据库。 引言中的计数来自完整的 magndata 列表，该列表于 2026-10-17 获取一次：公度条目 2,278 条，非公度条目 169 条（列表中的两个部分），共计 2,447 条。按约化学组成对条目分组，并剔除标签中未给出简单化学式的 169 条（其中大多数将群写在圆括号或方括号中；其余则为非化学计量或带有相名称），剩下 1,558 种材料，其中 464 种有两个或更多条目，这些材料共有 1,184 条目。

B 完整智能体详解

模型。 所有条件均在 `claude-sonnet-5` 上运行，每条回复最多包含 20,000 个词元，且不发送采样参数。评判者以及全文阅读器中的提取步骤均在 `claude-haiku-4-5` 上运行。

工具。

`search_paper` 会查询 OpenAlex、Semantic Scholar、Crossref 和 arXiv，合并并去重结果，剔除来自潜在掠夺性期刊的论文以及非英语论文，优先排列标题中首先提及该材料的论文，并在每次调用中返回最多 10 篇此前尚未展示过的论文，同时提供其元数据和摘要。连续展示的未提及该材料的论文最多为五篇。当一次查询找到的论文少于 3 篇时，

该工具会使用该化学式的其他拼写形式重复查询。当论文的标题或摘要以任何常见排版形式给出该材料的名称或化学式时，即视为该论文提及了该材料，包括带空格、下标或标记语言书写的化学式（“Ba_2IrO_4”、“$CoPS_3$”）。

`check_venue_status` 返回论文的出版类型、是否经过同行评审，以及是否已撤稿或来自可信度存疑的期刊。

`get_journal_quality` 返回来自 OpenAlex 的期刊级信号，并将 SCImago分区作为次要信号。

`read_paper` 仅从免费来源获取全文（OpenAlex 的开放获取副本、Semantic Scholar、设置访问密钥后可使用的 CORE、Unpaywall、存储库页面，最后才是 arXiv，以便优先选择正式出版版本而非预印本），将其缩减为最多 8,000 个字符的磁学相关内容；如果不存在免费全文，则返回摘要。

`compare_orders` 将附录 D 的匹配规则应用于两段描述，使用与评分器相同的代码和评判者，并由代码提供材料的晶系和点阵中心化信息。

`record_candidate_orders` 存储智能体找到的磁序；如果按照不含评判者且不考虑其所述条件的匹配规则，该磁序与已存储的任何磁序都不匹配，则该磁序为新的磁序。因此，在新条件（例如高压）下报告的同一磁序并不算新的磁序。

循环。 每种材料最多处理80篇论文、16轮、20次成功的全文读取和50次读取尝试。当已查看80篇论文后，搜索即停止（108次运行中有16次达到这一点），但读取和比较仍会继续。在108次完整智能体运行中，78次以模型给出答案结束，16次在连续三轮未出现新的磁序后结束，14次在提前停止时结束，0次达到16轮上限。记录了某种磁序后，凡是不增加新磁序的轮次，无论模型在其中进行了什么操作，都会计入停止规则。在记录任何磁序之前，只有进行过搜索的轮次才可能计入提前停止，而且该轮必须未找到任何提及该材料的新论文、未读取任何全文并且未记录任何内容；用于审查或比较的轮次不会改变计数。每经过上述任一种轮次后，模型都会获知在必须给出答案前还剩多少个这样的轮次。§2.1 的冲突轮次要求模型在答案中给出两个其标记为存在争议的磁序，并且代码判定这两个磁序不同，同时确认它们处于相同的明确条件下：二者均为实验性的，均针对标准结构形式的块体晶体，均处于环境压力下、无外场，并且都不局限于某个温度区间。随后，模型会获得一次搜索机会，该搜索必须旨在解决这项争议；还会获得一轮读取和比较机会，但不能搜索。如果在达到轮次上限前剩余轮次不足两轮，或运行已达到论文数量上限，则不会提供冲突轮次；这一情况在5次运行中发生。

答案核查。 代码会核查答案引用的每篇论文，确认其是否发表在被标记为可能不可靠的期刊上，以及其撤稿状态；在本次实验中，没有论文因这两项原因被排除。如果答案引用了智能体从未检索到的论文或被排除的论文，循环会将答案退回一次，列出相关引文并要求替换；本次实验中没有答案需要这样处理（108 次中为 0 次）。如果模型的最终回复中没有可读答案，循环会再请求一次；这种情况发生在 8 次运行中，且每次运行随后都给出了可读答案。如果再次失败，循环会保留模型在该次运行中给出的最后一份完整答案；如果没有，则会要求模型根据已收集的证据再作答一次。两种情况都未发生（运行次数分别为 0 和 0）。

置信度。 置信度评估标准按层级排列各项因素。首要因素是可重复性：依据匹配规则，评估独立研究团队报告相同序的情况是否一致，因此，单次且未经重复验证的报告不应获得高置信度。其次是发表日期和期刊质量；所属机构、资助、引用次数以及论文是否经过同行评审，则会在边际上调整分数。如果较新的研究在相同条件下修订了经过充分重复验证的序，则以修订版本为准。来自被标记为可能不可靠的期刊的论文以及撤稿论文一律排除。

C 运行智能体：输入与示例

该智能体通过 Python 调用， `run_agent` 位于代码仓库的 `agent/agent_loop.py`中。唯一必需的输入是材料的名称或化学式；其余各项都有默认值，除非**表 4** 另有说明，这些默认值与本文评估的智能体所用设置一致。语言模型也是输入项之一：用户只需指定 Claude 模型的名称，即可用任意 Claude 模型运行该智能体；更换模型不会改变提示词、工具或上限。无论该智能体使用何种模型， `read_paper` 中的评判者和提取步骤均在`claude-haiku-4-5`上运行。

表4： `run_agent`的输入。

输入	默认值	设置内容
`material` (文本)	必填	材料的名称或化学式，例如 `"CrI3"`。
`model` （文本）	`$ANTHROPIC_MODEL` 如果已设置， 否则为 `claude-sonnet-5`	智能体使用的语言模型：任意 Claude 模型名称。 整个运行过程使用同一个模型。
`tools` (列表)	六种工具，见**表格2**	智能体可以调用哪些工具。 `get_magndata_reference`， 它会查找 MAGNDATA 条目所引用的论文，可以 添加，但在本文的运行中处于关闭状态。
`claimed_orders` (列表)	none	要与文献核对的序（用于验证）。如果 没有指定，智能体会报告文献支持的内容，如本篇 论文 ·
`magndata_entry` (文本)	none	仅在以下情况下使用的 MAGNDATA 条目 ID： `get_magndata_reference`已启用。
`max_papers`， `max_rounds` （数字）	80，16	附录的论文和轮次上限B。
`criteria` （列表）	none	额外答案字段： `transition_temperature`中的任意一个， `temperature_range`、 `competing_states`、 `dimensionality`， `field_dependence`。该论文的运行要求 `transition_temperature`、 `temperature_range` 和 `dimensionality`，在每个条件下；它们不计分。
`disagreements` （是/否）	no	列出论文存在分歧的字段， 以及每种解读及其来源。在本文的运行中，每种 条件下均如此。
`compare_judges`	none	评判者会参考 `compare_orders` 。若没有参考内容，则采用 仅固定规则；本文的运行通过了评分器的评判者审核。
`budget`	none	用于统计本次运行令牌数的计数器。
`checkpoint_path` （文件）	none	每轮次结束后保存运行的位置，这样即使发生崩溃，最多也只会丢失 一轮次。
`top_p` （数字）	未发送	一种采样设置，仅由仍接受该设置的模型支持 （不是由 `claude-sonnet-5`）。
`verbose` （是/否）	no	打印每个轮次。

答案是一个包含三个部分的字典： `orders`，包含输出约定字段的磁序（附录 F）； `provenance`，记录所见的轮次、工具调用和论文，以及运行停止的原因；以及 `transcript`，与模型的完整交互记录。无需更改代码，也可以通过环境变量 `ANTHROPIC_MODEL` 或 `--model` 选项 `run_ablation.py`选择模型。示例输入：

```
from agent.agent_loop import run_agent

# claude-sonnet-5 and the six tools of this paper.
answer = run_agent("CrI3")

# The answer fields of this paper's runs.
answer = run_agent("CrI3", criteria=["transition_temperature",
"temperature_range"、 "dimensionality"]、disagreements=True)


# The same agent on another model.
```

```
answer = run_agent("CrI3", model="claude-opus-4-8")

# Another model, extra answer fields, and a checkpoint file.
answer = run_agent("NiI2", model="claude-haiku-4-5",
criteria=["transition_temperature"],
checkpoint_path="NiI2_checkpoint.json")

# Check a claimed order instead of discovering one.
answer = run_agent("MnTe", claimed_orders=[
    {"magnetic_order": "A-type antiferromagnetic",
     "propagation_vector": "(0, 0, 0)"}])

answer["orders"]       # the magnetic orders found
answer["provenance"]   # rounds、 tool calls、 papers、 why it stopped
```

以及对于评估集中的某种材料，从命令行运行：

```
python3 agent/eval/run_ablation.py --rungs S7 --materials CrI3 \
    --s7-tools full --model claude-opus-4-8 --repeats 1
```

D 完整的有序比较规则

每个条件都以文本形式给出规则，并在代码中实现一次，评分器和完整智能体的比较工具都会调用这段代码。代码还明确了文本留给判断的三个细节：温度容差、磁矩方向检查以及对掺杂样品的处理。当以下每一项都满足时，两种描述属于相同的序。该规则使用以下术语。

晶胞和设置。 晶体中重复出现的盒子。同一个晶格可以用多个晶胞描述，这些晶胞称为设置；晶胞也可能是有心 的（C、I、F 或 R），内部还包含额外的晶格点。
传播矢量 $\mathbf{k}$。 磁性模式从一个晶胞重复到下一个晶胞的方式，以所选晶胞的倒易晶格为单位表示；$\mathbf{k} = (0、0、0)$ 表示每个晶胞都具有相同的磁矩。
星形 $\mathbf{k}$。 由晶体对称性联系起来的$\mathbf{k}$ 向量。单$\mathbf{k}$ 磁序使用其中一个，多$\mathbf{k}$磁序则同时使用多个。

对称性标签。 磁空间群，即晶体及其磁矩的对称性，以 BNS 记号和目录编号表示；或不可约表示，用于标记磁矩如何在晶体对称性作用下变换。

有序类型和传播矢量。 它们指的是相同的有序类型，以及在同一晶胞中表示的相同传播矢量。在同一晶胞内，相差一个完整倒易晶格步长（有心晶胞允许的步长也包括在内）、整体符号相反，或在一个星形的对称等价矢量中选取不同矢量的向量，均视为相同。不同的调制类型（螺旋型、摆线型、纵向正弦型）代表不同的磁序。多个对称等价矢量同时有序（多$\mathbf{k}$）与单个矢量有序属于不同磁序；给出的矢量数量（两个与三个）也必须一致；答案未说明数量，则表示没有指定数量。如果一方未说明矢量，双方仅按类型匹配，除非双方都指定了特定模式：从不描述同一种磁序的两种模式（之字形与奈尔型）属于不同磁序，其他不同模式则交由裁决者判定。若答案未给出矢量而关键答案给出了，单凭类型匹配得一半分；若关键答案本身未给出矢量，则得满分（§2.3.3）。

晶胞。 传播矢量以所选单元格的步长计数，因此只能在同一单元格内比较向量。同一晶格的两种标准描述（菱方有心群的六方设置和菱方设置，或同一空间群的原胞和常规晶胞）属于同一晶胞；若一个晶胞不是另一个晶胞的设置、超晶胞，也不是不同结构多晶型的晶胞，则视为不同晶胞。当代码发现两个所述晶胞之间存在冲突，且晶格参数无法判定时，由晶胞判定器作出决定。未说明的晶胞记录为缺失，并跳过其检查。

条件。 仅在强外加场、高压下或掺杂样品中存在的序，与环境条件下的序不同；未提及外加场和压力意味着环境条件。当一方未说明向量，而另一方明确说明是单层、少层薄片或薄膜，与明确说明是块体晶体的情况相比，属于不同的序。转变温度是判断所指相的证据：无法在 2 K 或 10% 范围内调和的温度交由温度判定器处理，只有在这些温度描述的是不同相时，才会阻止匹配。未提及温度不会阻止匹配。

磁矩。 如果某一侧的磁矩为零而另一侧不为零，或者同一格点上的磁矩大小相差超过一半，则属于不同的磁序。一侧的磁矩位于基面内，而另一侧沿唯一轴排列，也属于不同的磁序；但旋转序（螺旋态）和立方材料除外，因为它们不存在唯一轴。所有磁矩在平面内进行刚性旋转，不会破坏匹配。

对称性标签。 如果双方给出的都是磁空间群，或都是不可约表示，则将同类标签进行比较，绝不能用标签替代有序类型。标签一致时，可以解决传播矢量缺失或不一致的问题，但不能解决晶胞冲突。只有当两个标签显然是同一标签的两种拼写时，代码才会确认它们相同；代码还会将目录（BNS）编号视为其所代表群的名称。任何存疑之处均交由裁决者处理。

仅由评分器应用。 在读取答案之后，评分器又添加了三项规定，这些规定不在模型看到的文本中。对于未说明传播矢量的铁磁序，将其视为$\mathbf{k} = (0,\ 0,\ 0)$；补全后的向量可以确认另一侧的向量，但绝不能与之矛盾，而且评判者始终会看到按原文写出的磁序。当同一材料的答案和关键磁序存在多种配对方式时，采用匹配磁序数量最多的配对；若数量相同，则优先采用被取代的关键顺序以及列表中最早的答案。只有当两种磁序都说明了传播矢量时，才会依据匹配规则和细节得分的晶胞层级来比较两个晶胞，因为没有向量就无法定义晶胞（附录 N）。

E 评判者全文

为何要对这些比较进行评判。 对于晶胞，比较晶系、点阵中心化和晶格参数的固定规则拒绝了正确的匹配，而更简单的检查——确认双方是否使用相同的空间群符号——也无法泛化：两种描述可能在不同符号下使用同一个晶胞，也可能共享同一个点阵中心化字母，但其中一个晶胞是另一个的超晶胞。对于温度，机械式修正方案（当双方都指明同类样品时忽略冲突）与忽略所有冲突得到的分数相同，因为出现冲突的关键磁序从未指明样品。代码无法判断约 45 K 的少层样品磁序与约 61 K 的块体晶体磁序属于不同体系；解读文本则可以。

构建。 每个评判者都在温度 0 下使用 `claude-haiku-4-5` ，并且只能看到其问题所涉及的字段。晶胞、温度和条件判定器以及裁决者通过多数票作出三次投票决定；引文判定器针对每条引文和主张各作出一次判定，并且只有在使用后才接受检查（见下文）。失败的调用保留固定规则的决定，且绝不会被保存。无法读取的回复仍然是一个答案：对于条件判定器，它保留关键词判定；对于其他评判者，它则算作反对匹配的一票。投票评判者的判定结果会被保存，键由评判者看到的内容、模型、投票数、问题和提示版本组成；引文判定器的判定结果则以引文、主张和提示版本为键保存，因此只有对完全相同的问题作出的判定才会被重新调用。

使用前验证。在读过早期实验后，我们才加入晶胞和温度判定器。在使用它们计算任何分数之前，我们先固定了在晶体学或样品信息足以作出判断时它们必须给出的答案：同一菱方中心晶格的不同设置属于同一晶胞；单斜晶胞与六方母相并非同一晶胞；任何一对文本中提及单层、少层或薄膜样品、不同多晶型或外加压力的情况，都描述了不同相。在该实验的70个不同问题中，48个有此类固定答案，而判定器在这48个问题上都给出了正确答案。我们发现一处不一致：Ba_2CeIrO_6 上两个几乎相同、存在争议的温度得到了相反的判定。条件判定器已

通过17个有记录答案的案例进行了检查，其中包括5个关键词误读句子的案例，并且17个答案均正确。裁决者通过123对已知答案的问题进行了检查（其中97对由自动方式构造，例如一个序与其自身比较，或与符号翻转或改变后的自身向量比较；6对来自早期实验的真实案例；20对则是采用不同记法的对称性标签），并答对了121对。两处错误中，一对在询问裁决者之前就被固定规则排除，另一处是一个带有不确定措辞的对称性标签，因此不再会交由裁决者处理。2026-10-18，我们用每次投票均重新发起调用的方式重复了全部四项检查，因为我们使用的模型网关一直返回缓存回复，导致最初的三次投票实际上只是重复了同一个答案。所有判定结果均与之前相同，而裁决者的123个问题中，没有一个问题的三次投票结果不一致（其他判定器只保存多数票，因此未测量其投票分歧情况）。

关于本实验自身的冲突。 报告分数所需的每项评判者判定结果都已保存，因此没有遗漏。本实验的答案引出了 123 个不同的晶胞与温度问题，其中一些问题在进入评判者环节前就已由较早的检查解决；对于每个答案固定的问题，评判者都给出了判定，只有两个例外（一个 $FePSe_3$ 晶胞和一个 $Na_2Co_2TeO_6$ 相），但这两个例外都不会改变分数。条件判定器和裁决者给出的判定结果也都从保存的判定结果中重放，没有遗漏。

引文判定器。 最初询问该判定器某篇摘要是否支持某项主张时，并未告知材料是什么；因此，它将一些主张标记为与摘要相矛盾，而那些摘要讨论的是另一种材料，或者把一项高压研究解读为反驳了一项常压主张。现在，判定器会获知材料是什么，这改变了许多判定结果，其中大多数转为“支持”。为了区分纠正错误与评判者放宽标准，我们根据判定器看到的内容，手动标注了 31 对判定结果，其中既有改变的，也有未改变的。最初编写的判定器在其中 19 对上判断错误，而我们采用的判定器在 8 对上判断错误，且错误方向不一：它认可了特定模式（例如之字形或 A型），但摘要只以更宽泛的方式描述了该模式；此外，它还有两次判定过于严格。在这类案例中，大约每四项判定就有一项错误，而“支持”这一判定对涉及特定模式的主张过于宽松。

F 三个条件及确保其可比性的规则

A 混杂因素在此指条件之间除所获证据之外的任何差异。该项目的规则实质上是：

> 条件之间唯一可以变化的因素是：（1）模型接收到的证据；（2）向模型告知的有关该证据的信息。模型、采样参数、答案令牌上限、输出约定（每个条件都必须填写的答案字段）、权衡证据的说明、匹配定义、置信度评估标准和评分器均保持一致。其中任何一项的差异都属于混杂因素，而不是条件差异。

在开发过程中，这条规则曾两次遭到违反，而且每次都在没有任何错误的情况下改变了经审慎评估得出的结论。第一次，要求完整智能体报告某个序适用的温度，而另外两个条件没有这一要求；由于评分器会在字段为空时跳过温度检查，省略该字段只会提高它们的分数。第二次，两个较简单的条件没有收到匹配定义或置信度评估标准，因此比较结果部分反映的是哪个条件获知了评分规则。如今，这条规则已在代码中强制执行：每个条件的提示词都由同一个文件构建，只有具备相应工具的条件才会收到工具说明；答案字段由一个函数统一定义；测试则会检查每个提示词的共用部分是否逐字完全一致、每个条件是否都收到匹配定义和置信度评估标准的所有等级，以及不具备工具的条件是否没有被告知要调用工具。只有完整智能体会收到评估标准中的一条规则，即排除可能属于掠夺性期刊的论文和撤稿论文，因为只有具备核查工具的条件才能据此采取行动。提示词太长，无法在此列出（闭卷版约 3,900 词，完整智能体版约 4,400 词）。每个条件的提示词均由代码仓库中的 `agent/loop_prompt.md` 和 `agent/prompt.md`生成；评判者的提示词则位于 `agent/scorer/adjudicator.py` （单元格、温度和裁决者）、 `agent/scorer/condition_judge.py` 和 `agent/eval/citation_faithfulness.py`。

任何条件下都不会缩短摘要，因为摘要最后一句往往会说明精细化序。单次搜索只发送一个查询，不附带完整智能体搜索所用的两个筛选条件（排除可能属于掠夺性期刊的论文和仅英文筛选条件），因此，来自可能属于掠夺性期刊的论文也可能进入其提示词；尽管两个较简单的条件都没有被告知这一排除条件，但只要引用此类论文，该序的引用分数就会被设为 0，所有条件均如此。

G 答案键来源

表格5 列出了答案键中的磁序及其来源论文。答案键中的每个值都与说明该值的段落所在文件及页码一并保存，这条记录也随代码一同发布。自动化测试会检查答案键中的所有值是否均未取自 magndata 条目文件。三种铁磁体的传播矢量 (0，0，0) 可根据定义得出，因此没有引用论文。Ba_2LaRuO_6,Ba_2LuReO_6 和第二种 NiI_2 磁序的传播矢量留空，因为没有来源以我们可用的形式给出这些值。La_2NiO_4,($\frac{1}{2}$，$\frac{1}{2}$，0) 的传播矢量为推断所得：其论文只给出了四方晶胞中的面内分量 ($\frac{1}{2}, \frac{1}{2}$)。我们收集的论文中，没有一篇给出 4 种材料的晶胞；另有 2 种磁序具有磁空间群。四种材料的传播矢量来自共同作者提供的论文，后来该作者确认了这些值。

H 分数定义与预注册

正确性。对于每种材料，所有通过匹配规则的报告磁序和关键答案中的磁序均构成候选配对，并选择匹配数最多的配对方式（附录 D）。匹配的磁序得分为 1，仅类型匹配得分为 0.5；如果关键答案给出了晶胞而匹配结果未说明晶胞，或其晶胞无法与关键答案中的晶胞进行比较，则得分为 0.75；当任一磁序未说明传播矢量时，不比较晶胞，得分为 1。综合所有材料，T 总得分与 N 关键答案中磁序的数量分别为，召回率为 T/N，精确率为 $T/$（$T + \sum c$）。惩罚值的总和遍历未匹配的报告磁序，每个磁序根据其给出的置信度 p增加惩罚值 c：$c = 0.6\,[1 - \cos(\pi p/160)]$，适用于 $p \le 80$ 以及上文中的 $c = 0.6 + 0.02\,(p - 80)$；如果模型将该磁序标记为有争议，则惩罚值减半；如果模型未提供置信度，则惩罚值为 1。若某个未匹配的磁序本应与某个关键答案中的磁序匹配，但该关键磁序已被另一个报告磁序匹配，则该未匹配磁序不增加惩罚值；如果它是在其他条件下描述该材料，且引文判定器发现有一篇被引用的摘要支持该描述（附录 N），同样不增加惩罚值。正确性为 $2PR/(P + R)$。此外还会计算并公布一个简单的 F_1，其中每个未匹配磁序均计为 1。

细节得分和排名第一。 细节得分是对所有匹配对的附录 I 中各级别取加权平均，权重分别为 10、8、4 和 3，分别对应类型、向量、单元格和对称性；关键答案未陈述的级别不计入，其余级别的权重则相应提高。若模型给出的最高置信度顺序与当前关键顺序匹配，则排名第一为 1，0 则为 0。

引文完整性。 DOI 可解析的引文得 1。0；若评审者认为摘要未涉及该主张，则降至 0。8。仅凭标题找到的论文得 0。5；若其摘要支持该主张，则升至 0。6，若摘要未涉及该主张，则降至 0。4。没有摘要的记录不予评审，因此该引文仍保留其存在所获的分数。找不到对应已索引论文的引文得 0。一个顺序的分数取其最佳引文的分数；但若引文指向一篇不存在的论文（DOI 无法解析，且标题与任何已索引论文都不匹配），或其摘要与主张相矛盾，则该顺序的分数设为 0；仅凭标题且未匹配到任何已索引论文的引文得 0，但不会因此将该顺序的分数设为 0。没有引文的顺序得 0；若某顺序的引文完全无法解析，则将其排除，而不是记为 0。

缺失部分。 当某次重复实验中某个部分的分数无法计算时，该部分的权重会分配给其他部分，而不是记为零分，因此无法匹配任何内容不会在正确性、排名第一和细节这三项上受到三重惩罚。

表 5： 答案键：每种序占一行，并列出其来源。来源是答案键针对该序的任何字段所引用的所有论文，因此一篇论文可能只支持一个字段，例如晶胞或转变温度。

材料	有序类型（状态）	**k**	来源
Ba_2CeIrO_6	A型反铁磁	(0, 0, 1)	[36, 37]
Ba_2IrO_4	基面反铁磁，磁矩沿着 [110]	$(\frac{1}{2}, -\frac{1}{2}, 0)$	[38, 39]
Ba_2LaRuO_6	IIIa型反铁磁	空	[40, 41]
Ba_2LuReO_6	I型反铁磁	空	[42]
Ba_2YReO_6	反铁磁	(0,0, 1)	[43]
$CoPS_3$	锯齿形反铁磁，小磁矩沿 c 的分量（电流） 共线锯齿形反铁磁，沿 a 的磁矩（已被取代）	(0, 1,0) (0, 1,0)	[44] [44]
$CrGeTe_3$	铁磁性	(0, 0, 0)	[45]
CrI_3	铁磁性	(0, 0, 0)	[46, 47]
CrSBr	A型反铁磁	$(0, 0, \frac{1}{2})$	[48]
$CrSiTe_3$	铁磁性	(0, 0, 0)	[45]
$FePS_3$	锯齿形反铁磁	$(0, 1, \frac{1}{2})$	[49–51]
$FePSe_3$	锯齿形反铁磁	$(\frac{1}{2}, 0, \frac{1}{2})$	[52, 53]
La_2NiO_4	双亚晶格奈尔反铁磁	$(\frac{1}{2}, \frac{1}{2}, 0)$，推断	[54]
MnTe	A型反铁磁	(0,0, 0)	[55, 56]
$Na_2Co_2TeO_6$	三重**k** 反铁磁（电流） 单-**k** 锯齿形反铁磁（已被取代）	$(\frac{1}{2}, 0, 0)$，三 arms $(\frac{1}{2}, 0, 0)$	[57, 58] [57, 58]
NiI_2	非公度螺旋磁结构，低于 60 K (电流) 共线反铁磁，60至73 K (电流)	(0。1384, 0, 1。457) 空	[59, 60] [59, 60]
UO_2	横向三重**k** 反铁磁 (电流) 单一-**k** 反铁磁（已被取代）	[0, 0, 1] [0, 0, 1]	[61] [61]
$YbBi_2IO_4$	条纹反铁磁	$(\frac{1}{2}, 0, 0)$	[62]

表6： 总分及其四个部分。

part	所衡量的内容	权重
正确性（置信度加权 F_1）	答案是否正确	45%
细节得分	细节是否足够充分，能够据此采取行动	25%
Top-1准确率	最有把握的序是否为当前序	15%
引文完整性	被引论文是否经得起检验	15%

预注册。 在早期版本的实验中完成前三种材料的运行后、进行其他运行之前，我们记录了权重和三条报告规则：总分绝不脱离其四个部分单独报告；无法计算的部分应分配其权重，而不是记为零分；不完整的评估应标记为不完整。在当时较小的数据集上，排名取决于权重（在约45%的权重组合下，闭卷和完整智能体分别排名第一）。我们还设计了另外两个部分，但未将其计入总分：其中一个部分会奖励报告已取代晶序，并给予较低的

表 7： 细节得分及其两个层级。只有在序匹配且数值正确时，该层级才得分；关键答案未说明的层级则跳过，而不记为零分。

条件	细节得分	k 层级	晶胞级
闭卷	0.721 ±0.020	0.872 ±0.076	1.000 ±0.000
单次搜索	0.411 ±0.044	0.417 ±0.049	0.283 ±0.157
完整智能体	0.730 ±0.047	0.822 ±0.062	0.798 ±0.094

总分中的权重：类型 10%，k 8%，单元格 4%，对称性 3%。对于关键答案没有值的级别，将跳过该级别，而不是计为零分。

置信度高于当前方法，因为合适的奖励取决于用户提问的原因；同时还包括校准部分，因为中等置信度下的粗略答案可能具有良好的校准效果，却只提供很少的细节。两者分别报告。总分是否与人们对答案的评判方式一致，尚未经过测试；计划中的检查是让一名评审者在看不到分数的情况下，对一组答案进行排名。

对权重的稳健性。 我们在所有由 5% 的整数倍构成且总和为 100% 的四个权重组合下重新计算了总分，共计 1771 种权重组合。在每一种组合下，完整智能体都位居第一，因为在四个部分中，它的平均值都是三个条件里最高的：正确性为 0.874、细节得分为 0.730、排名第一为 0.889、引文完整性为 0.867；相比之下，另外两个条件的对应值最高分别为 0.743、0.721、0.778 和 0.812。它相对于闭卷条件在细节得分上的领先幅度处于不确定性范围内（[-0.137，0.138]），因此在主要将权重放在细节上的权重组合下，两者并没有明显区分。预注册阶段在较小的早期数据集上发现的权重依赖性，在本次实验中并未出现。

添加一个标准。 这四个细节等级适合我们提出的问题，而其他问题则需要其他信息：转变温度、每个相的温度范围，或某种序属于体相还是单层。此处的标准是指在一个文件中共同声明的三项内容：要求智能体填写的字段、比较该字段两个值的规则，以及一个权重。声明了字段但未声明规则时，只报告该字段，不进行评分。每项标准都必须说明何时应将该字段留空，因为要求填写某个字段却不说明这一条件，会诱使人猜测一个值。在本次实验中，每个条件都报告了转变温度、各相的温度范围以及样品的维度，权重均为零。

I 该　　磁序词汇和细节评分等级

仅凭字符串无法判断两个有序类型名称哪个更具体，因此评分器使用了一个包含四个层级术语的词汇表：类别（铁磁性、反铁磁、亚铁磁性、非共线）、族（共线、共面、非公度）、主题（之字形、A型、奈尔、条纹）和方向。完整列表由评分器生成，并随代码一同发布。 **图 5** 展示了这些类别、族和命名自旋构型，以及它们的嵌套关系：每个命名自旋构型属于一个族，而一个族至多属于一个类别。

只有当匹配规则尚未通过传播矢量、调制类型或矢量数量判定某个术语时，该术语才能获得细节分数，因此这些内容不会被重复计分。之字形、A型和奈尔等特定模式可以获得分数，因为匹配规则检查的是传播矢量，而不是模式是否被明确命名。调制类型（螺旋型、摆线型、纵向正弦型）以及传播矢量的数量（单k对多k）不能获得分数，因为匹配规则已经对它们进行比较；磁矩方向也同样不能获得分数。此类术语也不能通过更宽泛的名称挽回其分数：如果因为“螺旋有序”属于旋转磁矩有序这一族而给予分数，就会把刚刚扣除的分数还回去，因此该族不获得细节分数。

有序类型、传播矢量和晶胞层级相对于关键答案分别获得 1、0.5 或 0 分。对于比关键答案恰好低一个精确层级且与其一致的答案，给予一半分数：例如，关键答案指定了一个矢量时，答案给出全部对称等价矢量；关键答案还给出晶格参数时，答案给出空间群符号；或者关键答案指定了诸如之字形这样的模式时，答案给出正确的族，例如“共线”。仅有“反铁磁”这样的类别，只有在关键答案中最细的术语是倾斜型时，才能获得一半分数，



图5： 评分器定义的细节得分所采用的有序类型词汇。每条连线都将一个术语与其所属的上位术语相连；共线、公度和非公度均不表示任何类别。除非方括号中注明类别，否则所有命名图案均为反铁磁，且这些调制均为非共线。拓扑织构包括斯格明子及相关织构；IIIa型和IIIb型是III型的更细分类。带†标记的术语不计细节分，其中大多数是因为匹配规则已经对它们进行比较；未显示的传播矢量数量和磁矩方向相关术语也不计分。

表 8： 按条件划分的引文完整性。“跳过”表示其引文均无法解析的磁序；这些磁序会被排除，不计为 0 分。

条件	引文完整性	已检查的订单	跳过
闭卷	0.425 ± 0.102	151	7
单次搜索	0.812 ± 0.035	372	0
完整智能体	0.867 ± 0.020	229	0

已应用的检查：完整检查（论文确实存在、未被排除，且其摘要支持该主张）。跳过的检查不计入平均值，而不是记为失败。引文完整性先按每次重复实验中的磁序取平均值，再对重复实验取平均值，与总分的计算方式相同；若直接对所有磁序取平均值，结果会略有不同（此处闭卷为 0.4253，合并计算为 0.4265）。

共面或非共面族（均属于非共线类）；若与一个命名图案的关键答案，或与共线、公度或非公度族进行比较，则不得分。与关键答案相矛盾的图案（关键答案为 A型 时却为 G型）得零分。对称性级别与关键答案的标签一致时得 1 分；当一方列出包含另一方单一标签的备选项时得 0.5 分；其他情况得 0 分。在本实验中，关键答案为七种磁序规定了对称性标签：Ba_2IrO_4，Ba_2YReO_6,La_2NiO_4 和两种 $CoPS_3$ 磁序使用不可约表示，$YbBi_2IO_4$ 和 $Na_2Co_2TeO_6$的已取代晶序则使用磁空间群。

J 完整的引文核查

第一步分别核查引文的标识符是否能够解析，以及论文是否存在，因此无法解析的标识符不会被计为伪造。搜索分别通过 DOI 和标题进行，因为在闭卷条件下，引文通常只提供论文标题，或作者和年份，而没有 DOI。标题先在 Crossref 中查找，再在 OpenAlex 中查找，因为 Crossref 对较早文献的索引不完整。每次查找都会保存，并归档在所查找的 DOI 或标题下，因此重新评分时无需网络访问。第二步针对每一条已找到论文（通过 DOI 或标题找到的引文），根据该论文书目记录中的摘要进行评审。在本实验中，答案给出了 589 条不同的引文。其中 41 条仅通过标题提供，其最接近的索引匹配既相似到无法排除，又不同到无法接受，因此不予评分，而是直接排除。书目记录中没有摘要的论文不接受评审，但会保留其已存在这一项的得分。

K 可复现性

表格以及从生成的数字文件中提取的所有数字，均由代码仓库中的脚本根据保存的答案写入；每位评审者的判定结果也都根据保存的判定结果重放，因此重新生成这些内容无需网络访问，也无需模型调用。智能体本身的答案无法通过再次运行来复现，因为模型并非确定性的。闭卷答案来自较早的一次实验；我们核实了从系统提示词开始，该条件发出的每个请求在那次实验和本次实验的代码下完全相同，因此沿用了这些答案，只重新运行了单次搜索和完整智能体。本次实验没有任何运行被搁置或重新运行，且每次运行都使用同一版本的代码。我们报告的标准差是根据重复试验总分计算的；一种材料的某次重复试验与另一种材料的哪次重复试验分组是任意的，因此它只能作为粗略参考，而非置信区间。开发期间，我们使用当时相同的检索方式和评分器，在三种材料上重新运行单次搜索，其平均正确率从 0.685 变为 0.643，变化幅度超过了各次运行内的标准差（0.031 和 0.021），因此我们将低于约 0.05 的差异视为噪声。单次搜索和完整智能体运行共使用约 5300 万个模型词元；在生成这些数字的修订版本中，智能体的离线测试套件（Python 3.9 下的 2,475 项测试）全部通过。

置信区间。结果中的区间是 95% 百分位数自助法区间。每个 10,000 次重抽样样本都会有放回地抽取 18 种材料，保留每种材料的全部 6 次重复实验，并按照论文中的计算方式，使用相同的评分器和相同的已保存评审判定，重新计算每个数值：总分、其四个部分、每次重复实验的过时度分数和校准误差，然后再计算重复实验的平均值，以及

表9： 结果中的数值及其95%置信区间，基于对18种材料的重抽样得出。(a) 各条件。(b) 条件之间的差异。

(a)	闭卷	单次搜索	完整智能体
总分	0.548 [0.429, 0.655]	0.676 [0.607，0.740]	0.839 [0.798，0.881]
正确性	0.536 [0.381，0.662]	0.743 [0.664，0.812]	0.874 [0.822，0.923]
细节得分	0.721 [0.612，0.830]	0.411 [0.289，0.539]	0.730 [0.621，0.831]
排名第一	0.417 [0.222，0.611]	0.778 [0.602、0.926]	0.889 [0.768, 0.981]
引文完整性	0.425 [0.281, 0.564]	0.812 [0.754, 0.861]	0.867 [0.825、0.909]
无传播向量的序	32% [19%，47%]	85% [75%，93%]	50% [35%，67%]
其引用的标题无法对应到任何论文	25% [14%，39%]	2% [0%，6%]	0% [0%，0%]
与……相矛盾的引用摘要主张	17% [9%，26%]	5% [4%，6%]	7% [4%，9%]
电流与已被取代的（过时度分数）	0.33 [0.00，0.67]	0.69 [0.08，1.00]	0.64 [0.42、0.83]
期望校准误差	0.266 [0.196, 0.391]	0.235 [0.165、0.316]	0.135 [0.126，0.239]
平均置信度，置信度订单 ≥ 80	87 [86, 88]	88 [87, 89]	89 [88, 90]
与关键答案匹配的置信度订单 ≥ 80关键答案	65% [41%，87%]	76% [59%，98%]	85% [70%，98%]

(b)	总分	正确性
单次搜索相对于闭卷	0.128 [0.016、0.255]	0.207 [0.082, 0.362]
基于单次搜索的完整智能体	0.163 [0.102，0.229]	0.131 [0.053，0.217]
基于闭卷的完整智能体	0.291 [0.171，0.427]	0.338 [0.194，0.511]

（b，续）	单次搜索相对于闭卷	基于单次搜索的完整智能体	基于闭卷的完整智能体
细节得分	[-0.473, -0.137]	[0.205，0.428]	[-0.137, 0.138]
排名第一	[0.139，0.593]	[0.009，0.250]	[0.287，0.667]
引文完整性	[0.236，0.539]	[-0.014, 0.139]	[0.283, 0.610]
无传播向量的序	[33%，69%]	[-47%，-22%]	[-4%，38%]
与……相矛盾的引用摘要主张	[-21%，-4%]	[-1%，4%]	[-20%，-2%]
期望校准误差	[-0.155，0.053]	[-0.121, 0.019]	[-0.211, -0.019]
置信度订单 ≥ 80 匹配关键答案	[-9%，43%]	[-3%，23%]	[-2%，51%]

所有重复实验合并计算的百分比和高置信度磁序的平均置信度。我们对材料而不是重复实验进行重抽样，因为材料之间的差异远大于重复实验之间的差异，因此区间显示了在另一组材料上某个数值可能发生多大变化。三个条件使用相同的重抽样样本，因此两个条件之间差异的区间是配对区间。完整智能体的期望校准误差接近其区间的下限：该误差累加置信度与准确率之间的绝对差距，因此重抽样引入的噪声往往会使其升高。过时度分数仅适用于具有已取代晶序的 3 种材料，因此其区间取自至少包含其中一种材料的重抽样样本。 **图 4**则改为对每种材料的 6 次重复实验进行重抽样，因为单个材料没有其他可供重抽样的对象。

表 9 给出了结果中每个具有区间的数字的区间。

L 逐材料结果及匹配规则审计

图 4 报告的是正确性，而非按材料统计的总分，因为对于每种材料，只有当序匹配时才会有细节得分，因此按材料统计的总分会将不同材料的不同部分合并在一起。

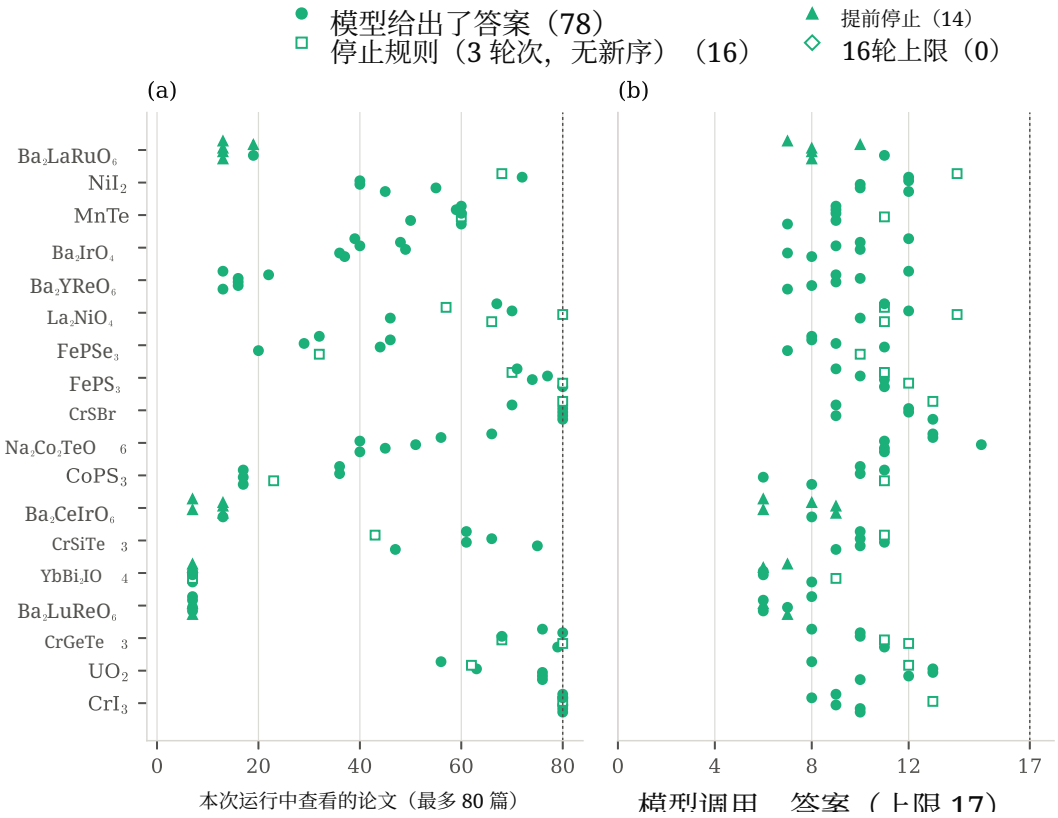

图6： 每次完整代理运行各用一个标记表示（每种材料6个），标记形状显示了运行结束的原因。(a) 运行搜索返回的论文。(b) 运行进行的模型调用次数，包括给出答案的那次调用，因此因16轮上限而结束的运行显示为17。材料顺序与**图4**相同。

完整智能体的运行如何结束。 搜索找到的文献大致将材料分为两组（**图 6**）。对于 Ba_2CeIrO_6, Ba_2LaRuO_6, Ba_2LuReO_6, Ba_2YReO_6 和 $YbBi_2IO_4$, 没有任何一次运行找到超过 22 篇论文，所有 14 次提前停止都出现在这些材料上。在其余 13 种材料中，每次运行至少找到 36 篇论文，$CoPS_3$ 和 $FePSe_3$ 除外（17 至 46 篇）；108 次运行中有 16 次达到了 80篇论文上限。此后，智能体继续解读并比较，直到停止规则、轮次上限或其自身给出的答案结束运行。

零分的分布情况。闭卷模式下，得分恰为零的49个答案分布在18种材料中的13种上；对于Ba_2CeIrO_6, $FePSe_3$ 和$YbBi_2IO_4$，每次重复都得零分（**图7**）。单次搜索的7个零分答案分布在2种材料上，其中6个来自Ba_2LaRuO_6, ；完整智能体的1个零分答案来自Ba_2LaRuO_6。闭卷模式的6次弃答全部来自$YbBi_2IO_4$, ，单次搜索的6次弃答则全部来自Ba_2LaRuO_6。对于$FePS_3$ 和$FePSe_3$, ，闭卷模式在每次重复中都给出锯齿形反铁磁序，且给出的传播矢量均为(0，1，0)，置信度为55至90；而关键答案分别为(0,– 1，$_{1\,2}$)和($\frac{1}{2}, 0, \frac{1}{2}$)。

三个材料揭示了停止规则或关键答案的影响。§2.1 中的提前停止有利有弊，而这两方面在单个材料上都能体现出来。它在4种材料的14次运行中触发。在 Ba_2LaRuO_6 上，它结束了6次运行中的5次：搜索几乎没有找到关于该材料的论文（每次运行最多找到19篇），其中最早的两篇来自20世纪80年代初，其中一篇是我们关键答案的来源 [40], ，因此继续搜索几乎找不到更多内容。在 Ba_2CeIrO_6 上，它结束了6次运行中的5次，而且往往在确定序的中子研究 [36] 出现之前就停止了。这项研究没有被任何一个完整智能体答案引用；而在先前一次使用未设置提前停止的智能体版本进行的实验中，6个答案里有5个引用了这项研究，该版本通过未提及材料名称的搜索找到了它。在 Ba_2YReO_6 上，完整智能体在6次重复实验中有5次未能排名第一，因为它最有把握的序判断引用了一项2025年研究，该研究将奈尔温度定为31K，并指出其上方存在四极相 [63], ；而我们取自较早一项中子研究的关键答案 [43], 记录了在45 和20K时发生的相变；温度判定器将二者视为不同相。这里

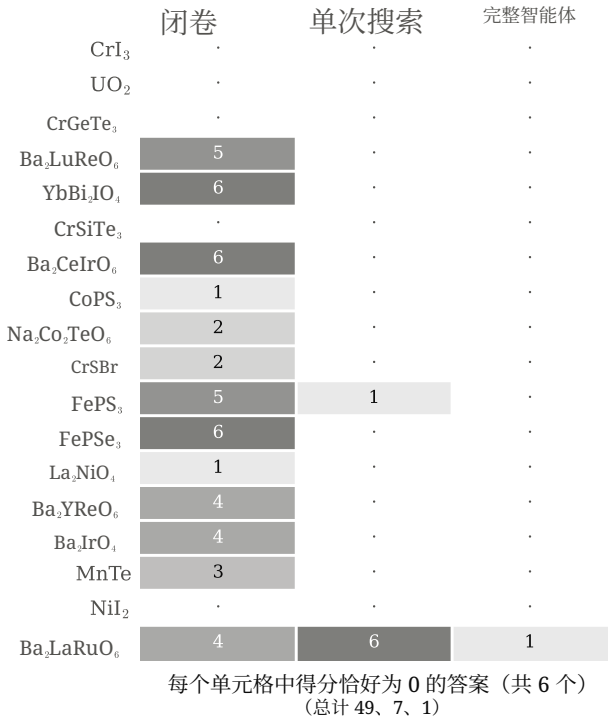

图7： 每种材料和条件下6个答案中分数恰好为零的答案；圆点表示没有。材料按 **图 4** 的顺序排列。

智能体遵循的文献比我们的关键答案更新。

逐种材料比较当前磁序与已取代晶序。§3.4中的过时度分数为：当当前磁序获得的置信度高于已取代磁序时，分数为1；仅报告当前磁序时，分数为0.5。在UO_2 中，每个条件的每次重复实验得分至少为0.5；单次搜索在全部6次重复实验中均得分1，完整智能体在4次中得分1，闭卷在2次中得分1。在$CoPS_3$ 中，完整智能体在两次重复实验中得分1，其余四次得分0.5；闭卷在四次重复实验中得分0.5，其余两次得分0；单次搜索在一次重复实验中得分0.5，其余均得分0。在$Na_2Co_2TeO_6$ 中，单次搜索在每次重复实验中均得分1，完整智能体在两次中得分1（另有一次得分0.5）；其他每次重复实验均得分0。在该材料中，当前三重-k描述和已取代的单-k描述在单次搜索和完整智能体的重复实验中，除各有一次外，均同时被报告并标记为存在争议；这既反映了文献中的持续争议，也同样可能是错误所致；闭卷从未报告当前磁序。

未匹配磁序的含义。图8 按每项报告的磁序所描述的内容进行分类。每项报告的磁序都会按照以下规则中第一个适用的规则打上标签：磁序或其条件中提及的理论方法；非块体样品；或掺杂、高压或外加场。若这些规则均不适用，则计为常压块体晶体。在未匹配的磁序中，有 11、50 和 7 项（分别对应闭卷、单次搜索和完整智能体）重述了同一答案中另一项已匹配的磁序所描述的内容；另有 27、192 和 69 项描述的并非常压块体晶体；一项磁序可能同时属于这两类。完整智能体在 18 种材料中的 14 种上报告了常压块体晶体以外的磁序。应用附录N中的规则后，所有材料和重复实验中剩余的惩罚值总和分别为 28.3、4.9 和 13.0。每项磁序、其标签及惩罚值均与论文的来源一并公布。

逐项审计。依次关闭每项可能直接否决匹配的检查，并关闭评判者，对报告顺序与关键顺序的全部975次比较重新评分（表格10a）。如果没有固定晶胞规则，总分将分别上升 $+0.005$、 $+0.000$和 $+0.001$；如果没有温度规则，则分别上升 $+0.002$、 $+0.000$和 $+0.012$；这两次比较现已提交给评判者。图版（b）显示了没有评判者、加入任一评判者或同时加入两位评判者时的分数。

表10： 匹配规则的各个部分如何影响分数。

（*a*）逐一关闭每条拒绝匹配的规则

规则拒绝于	它首先拒绝的配对	闭卷	单次搜索	完整智能体
磁矩方向	24	+0.017	-0.001	-0.008
晶胞	9	+0.005	+0.000	+0.001
温度	40	+0.002	+0.000	+0.012
阶数类型	255	+0.002	+0.023	+0.005
单 k 与多 k	80	+0.000	+0.000	+0.000
k 臂数量	2	+0.000	+0.000	+0.000
磁矩大小	0	+0.000	+0.000	+0.000
外加场 / 压力	93	+0.000	+0.001	+0.002

(b) 将晶胞与温度问题发送给评审者，可以发送其中一个或另一个，也可以两者都发送

发送给评审者的问题	分数的一部分	闭卷	单次搜索	完整智能体
两者皆非（仅固定规则）	总分	0.543	0.676	0.833
	正确性	0.530	0.743	0.863
	细节	0.717	0.411	0.733
	排名第一	0.407	0.778	0.870
仅晶胞	总分	0.546	0.676	0.833
	正确性	0.532	0.743	0.863
	细节	0.724	0.411	0.733
	排名第一	0.417	0.778	0.870
仅温度	总分	0.544	0.676	0.839
	正确性	0.534	0.743	0.874
	细节	0.714	0.411	0.730
	排名第一	0.407	0.778	0.889
两者（报告值）	总分	**0.548**	**0.676**	**0.839**
	正确性	0.536	0.743	0.874
	细节	0.721	0.411	0.730
	排名第一	0.417	0.778	0.889

面板（a）：在两位评判者均关闭的情况下，对于报告顺序与关键顺序之间的全部975次比较，关闭相应规则时每个条件的总分变化了多少。正值表示固定规则拒绝了原本可以得分的答案；负值表示它拒绝了本不应得分的答案。粗体显示的两条规则，就是现在提交给评判者的规则。图版（b）：将晶胞问题、温度问题或两者提交给评判者，而不是由固定规则判定时的总分及其构成部分。

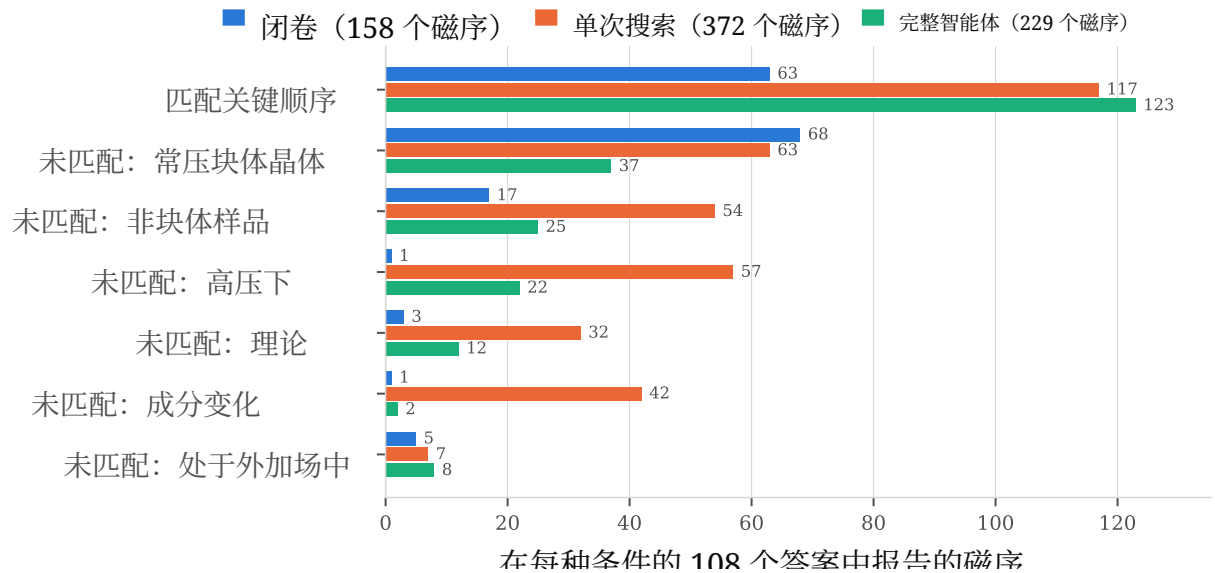

图8： 每种条件下报告的所有序，按其答案统计：其中哪些与关键顺序匹配，哪些与任何关键顺序都不匹配，并按其描述的内容分类。在同一答案中，若某个未匹配的序复述了一个已匹配的序，则将其计入其所描述的类别。

M 细节方面的局限

评分器。细节得分会奖励给出正确的数值，但只有当某个磁序明确说明了温度、磁场或压力时，与之冲突的条件才会导致匹配被否决。因此，按照这些标准，评分器倾向于给未注明条件的答案更高评价。关键词决定条件判定器会读取哪些磁序，因此，措辞中没有任何关键词能识别的磁序，会被视为处于环境条件下，而不会经过评审者判读。

比较所揭示的差异。我们没有记录每种材料对应的文献数量，因此无法判断这能在多大程度上解释闭卷方法在不同材料上的表现差异。对同一次运行进行重复实验，其结果的变化也小于不同运行之间的变化（附录 K）。

N 实验后的评分决策

答案读取完毕后，我们修改了评分器；每次修改都同样适用于所有条件，并针对已保存的答案进行，因此没有重新运行任何实验。大多数修改都是为了纠正评分器对答案内容的读取错误：例如，它误读了传播矢量的记法、对称性标签或晶胞；把“无长程有序”这样的否定表述当作主张；或者漏掉了较长分句中提到的磁矩方向。每项修改都是通过阅读答案发现的，而每项修改所影响的所有答案也都经过了阅读。另有几项修改关乎评分应当奖励什么，而非纠正错误，其中最主要的包括：将铁磁体未填写的传播矢量视为 $(0,0,0)$；采用与最多磁序相匹配的配对方式，而非先到先得；仅当两个磁序都注明传播矢量时才比较两个晶胞；以及规定，当未匹配的磁序重复了某个已匹配磁序的内容，或描述了关键答案未涵盖条件下的材料，且有引文摘要支持时，不予扣分。我们记录了最后一条规则，并在计算其影响之前就确定了该规则。还有一项修改涉及引文判定器：现在会告知它主张所涉及的材料，因此，关于相关材料的论文将不再能为该主张提供支持。最后五项修改是在读完本次实验的完整智能体答案后决定的。表格11 展示了随着这些修改逐项应用，总分和正确性的变化。最后一行是报告的结果；第一行则是完全不作修改时的分数，而各条件的排名顺序已经与最终结果相同。完整智能体是在除最后五项之外的修改全部完成后运行的，其比较工具使用的也是该评分器，因此对于这一条件，第一行显示的是反事实分数，而非当时实际给出的分数。

单元格和温度　　在早期实验完成后，也将文献评审者纳入了匹配规则　　n

表 11： 随着在阅读答案后所作的评分调整逐项累积应用，总分（括号内为正确性分数）。一行列出多项调整时，给出的是完成整组调整后的分数；代码仓库的RESULTS.md 文件将每项调整分别列在单独一行中。

调整（累积）	闭卷	单次搜索	完整智能体
在未应用以下任何调整的情况下评分	0.529 (0.560)	0.567 (0.519)	0.712 (0.695)
+ 与最多磁序匹配的配对（此前采用先到先得）	0.528 (0.568)	0.567 (0.519)	0.721（0.708）
+ 铁磁体的空白 k 读作（0，0，0）	0.528（0.568）	0.606（0.569）	0.748（0.739）
+ 4 进一步修正评分器读取答案的方式	0.528 (0.568)	0.609 (0.574)	0.780 (0.767)
+ 重述，并在其他条件下支持磁序，不花费任何东西	0.536 (0.585)	0.625 (0.610)	0.783 (0.774)
+ 23评分器读取答案方式的进一步修正	0.533 (0.541)	0.624 (0.587)	0.802 (0.789)
+ 告知引文判定器该主张涉及哪种材料	0.543 (0.541)	0.652 (0.698)	0.800 (0.829)
+ 5 进一步改进评分器读取答案的方式	0.536 (0.531)	0.652 (0.698)	0.812 (0.839)
+ 仅当两种磁序都给出传播向量	0.543 (0.531)	0.663 (0.723)	0.825（0.851）
+ 7 进一步修复评分器读取答案的方式	0.546（0.531）	0.659（0.716）	0.825（0.851）
+ 根据本次实验的完整智能体答案确定的五项更改已读取	0.548（0.536）	0.676（0.743）	0.839（0.874）

读取；若不计入这些内容，总数将按相同顺序为 0.543、0.676 和 0.833。

O 在其他智能体模型上进行相同的实验

为了了解比较结果是否取决于智能体模型，我们在另外三个模型上进行了相同的实验： `claude-opus-5-5`、`claude-fable-5-1` 和 `claude-opus-5`。四个模型都在相同条件下参与了同一项实验：对每个模型都进行了单次搜索和完整智能体运行，使用相同的代码、已保存的评审判定和论文摘要，因此任何模型的运行都不会使用其他模型的结果；闭卷答案则重复使用，因为相关请求完全相同，且不涉及文献。对于另外几个无法访问 arXiv 的模型，我们使用相同的智能体代码重新进行了七次完整智能体重复实验。唯一不同的是智能体模型；提示词、工具、上限、答案键、评分器和评判者（`claude-haiku-4-5`）均相同。 `claude-sonnet-5` 进行了 6 次重复实验，这是本文的主要结果；其他每个模型均进行了 3 次，因此其标准差基于较少的重复实验（**表12**，**图9**）。

完整智能体在每个模型上的总分平均值都最高，并且对单次搜索的领先优势在每个模型上都成立：如附录K所示，对材料进行重抽样后，差值的95%区间不包含零。只有对于 `claude-sonnet-5` 和 `claude-opus-5`，它对闭卷的领先优势才明确；对于 `claude-opus-5-5`和 `claude-fable-5-1`，区间则包含零。对于这两个模型，闭卷的平均分也高于单次搜索，因为在`claude-sonnet-5` 未能给出关键答案的传播矢量的材料中，这些模型能凭记忆给出该矢量（例如 $FePS_3$ 和 $FePSe_3$），但这一差异也处于不确定性范围内。**图10** 展示了另外三个模型各自的四个部分：与 `claude-sonnet-5`一样，对于每个模型，单次搜索在三个条件中提供的细节最少。

表 12： 每个智能体模型在各条件下的总分，取重复实验的平均值 ± 重复实验间的一个标准差；评分使用与主要结果相同的评分器，并保留评审者判定。

智能体模型	重复实验	闭卷	单次搜索	完整智能体
十四行诗 5（主要结果）	6	0.548 ± 0.044	0.676 ± 0.031	0.839 ± 0.027
Opus 5.5	3	0.805 ± 0.009	0.710 ± 0.007	0.831 ± 0.020
Fable 5.1	3	0.773 ± 0.022	0.732 ± 0.003	0.828 ± 0.029
Opus 5	3	0.679 ± 0.015	0.701 ± 0.029	0.853 ± 0.024

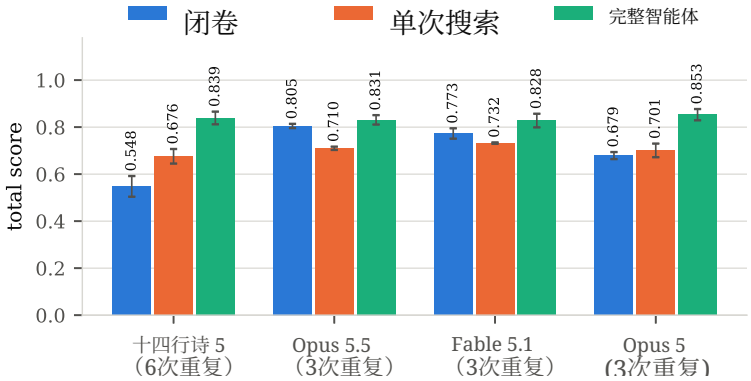

图 9: 的总分 **表格 12**；误差线表示重复实验间的一个标准差。

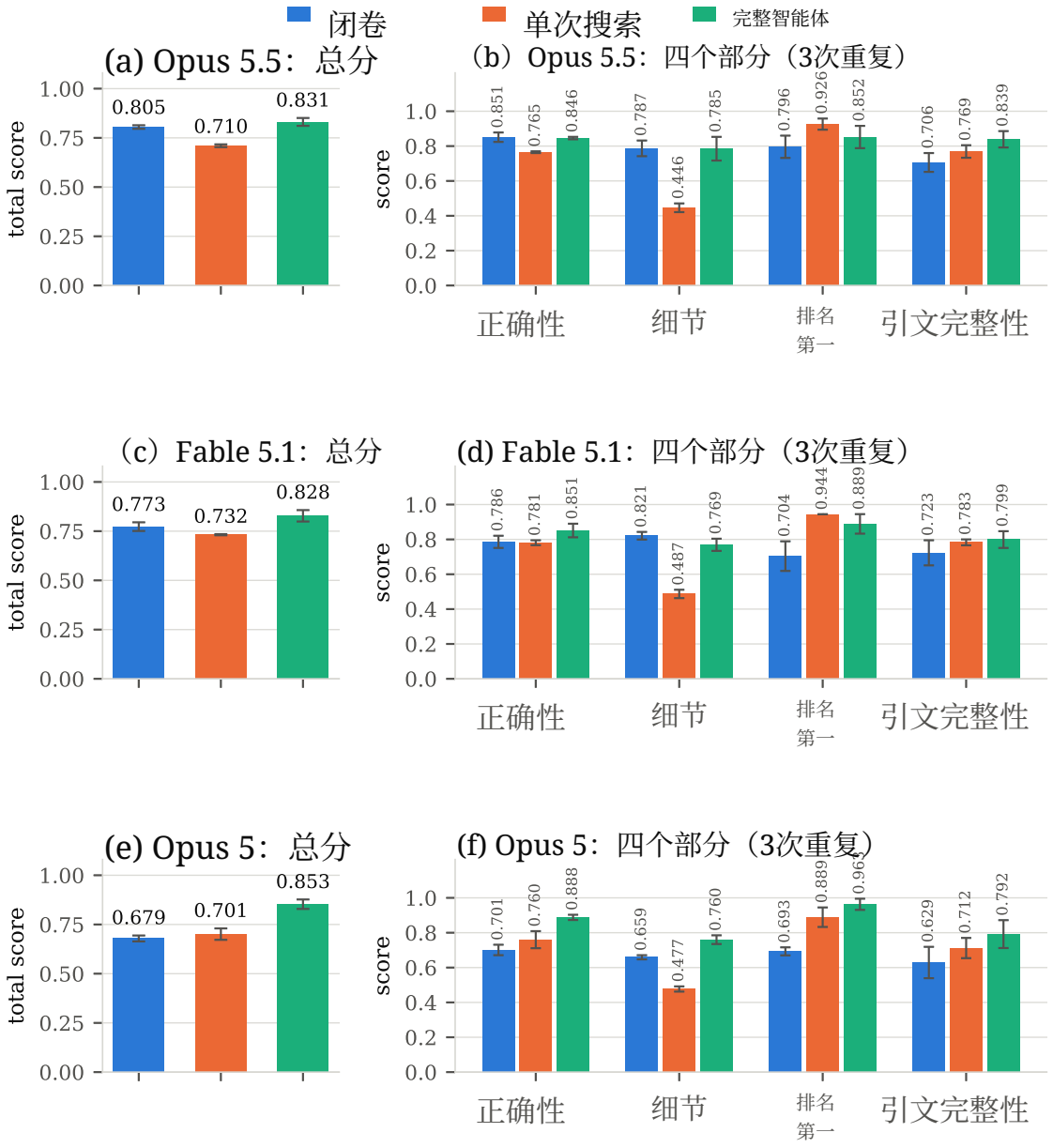

图10： 图 2a 和 b 展示了另外三个智能体模型：对于每个模型，分别给出总分（左）及其四个组成部分（右），数据为 3 次重复的平均值，误差线表示重复实验间的一个标准差。35