SIN-Bench：追踪长上下文多模态科学交错文献中的原生证据链

Yiming Ren[1], [2*] **Junjie Wang**[1], [3*] **Yuxin Meng**[1*] **Yihang Shi**[1*] **Zhiqiang Lin**[1] **Ruihang Chu**[1] **Yiran Xu**[1] **Ziming Li**[4] **Yunfei Zhao**[3,5] **Zihan Wang**[3,6] **Yu Qiao**[2†] **Ruiming Tang**[4] **Minghao Liu**[3] **Yujiu Yang**[1†]

[1]清华大学 [2]上海人工智能实验室 [3]2077AI [4]快手科技 [5]斯坦福大学 [6]哈佛大学

rym24@mails.tsinghua.edu.cn, wangjunjie@sz.tsinghua.edu.cn, qiaoyu@pjlab.org.cn, yang.yujiu@sz.tsinghua.edu.cn

https://github.com/IIGROUP/sin-bench

arXiv:2601.10108v1 [cs.CL] 15 Jan 2026

摘要

评估多模态大语言模型是否真正理解长篇科学论文仍然颇具挑战：仅依据答案评分的指标和合成的“大海捞针”测试往往只奖励答案匹配，却不要求模型在文档中给出因果性、与证据关联的推理链。我们提出“海中捞鱼”（FITO）范式，要求模型在原生科学文档中构建明确的跨模态证据链。为使 FITO 能够落地，我们构建了 SIN-Data，这是一个保留文本和图表原生交错形式的科学交错语料库。在此基础上，我们构建了 SIN-Bench，包含四项循序渐进的任务，涵盖证据发现（SIN-Find）、假设验证（SIN-Verify）、有依据的问答（SIN-QA）和以证据为锚点的摘要合成（SIN-Summary）。我们进一步提出“无证据，不评分”原则：仅对有可验证锚点支撑的预测进行评分，并通过匹配度、相关性与逻辑性诊断证据质量。对八种多模态大语言模型的实验表明，证据定位是主要瓶颈：Gemini-3-pro 的平均总体得分最高（0.566），而 GPT-5 的 SIN-QA 答案准确率最高（0.767），但其证据对齐总体得分较低，暴露出答案正确与具备可追溯支撑之间的差距。

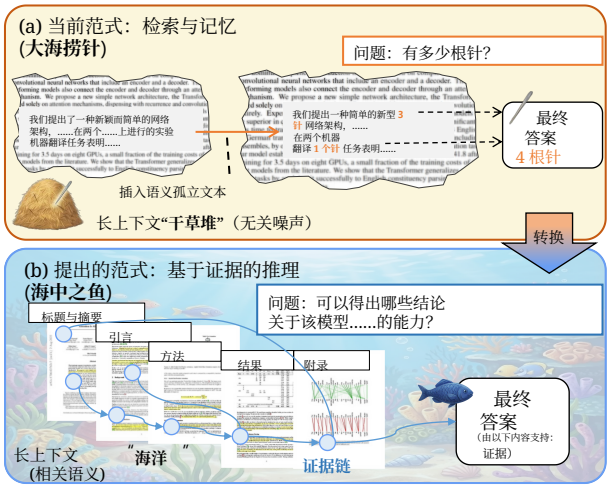


图 1：长上下文多模态评测范式对比。(a) 当前的 NIAH 方法将人工“针”嵌入无关噪声中，侧重于表层检索。(b) 所提出的 FITO 范式在原生文档生态系统（“海洋”）中评估深层理解。它要求模型跨章节聚合互联知识单元（“鱼”），形成用于推理的证据链。

1 引言

对长篇多模态科学文献进行深入分析，是人类认知的一大标志（Xia 等人，2024年；Chen 等人，2024年；Bai等人，2025a）。因此，这项任务是评估专家级推理能力的重要基准，例如博士研究人员所具备的专业水平。多模态大型语言模型（MLLMs）的进步（Bai 等人，2025b；OpenAI，2025年；DeepMind，2025年；xAI，2025年），尤其是在长上下文处理和多模态对齐方面的进展，使实现这种专家级阅读能力成为可能。然而，当模型处理这些篇幅冗长且符号稠密的文档时，现有评估体系无法区分系统究竟真正理解了复杂的文档逻辑，还是仅仅依赖参数化知识来推断答案。

尽管长上下文评估取得了快速进展，但现有基准测试与真实科学认知之间仍存在显著错位（Fu 等人，2024年）。这种差异主要体现在任务构建和评估指标方面。首先，就**任务设计**而言，如图1 （a）所示，当前普遍采用的“大海捞针”（NIAH）范式倾向于插入人工构造且语义孤立的片段（Wang

*贡献相同。
†通讯作者。审稿中。

等人，2024b）。尽管这种合成配置能够有效测试检索边界，但无法模拟真实文献所固有的天然性和长程依赖。例如，对图的解读通常需要回溯参考方法部分中的实验设置。其次，就**评估指标**而言，完全依赖答案准确率会掩盖推理过程的不透明性（Mathew 等人，2021年；Xia 等人，2024年）。这种方法使模型能够利用参数化知识走捷径，当系统面对科学问题时，便会产生幻觉。

为弥合这一差距，我们主张从仅关注答案正确性，转向关注跨模态证据链的构建，从根本上改变评估重点。因此，如图1 （b）所示，我们提出“海中之鱼”（FITO）评估范式。为证明模型并非仅凭猜测作答，系统必须明确阐述证据检索的过程（Dasigi 等人，2021年）。不同于 NIAH 中人工植入的“针”（噪声中的人工事实），“海洋”（科学文献）中的关键信息更像是原生的“鱼”（彼此关联的知识）。因此，解答科学问题本质上需要捕捉并关联分散在摘要、图和附录等组成部分中的这些“鱼”。模型的核心能力在于从广阔的信息空间中识别高价值证据，并将这些发现整合成自洽的逻辑链。

为将海中之鱼（FITO）范式付诸实践，我们构建了 SIN-Data，这是一套专为长篇科学文档设计的统一**科**学**交**错数据基础设施。我们汇集来自 arXiv、PMC 等来源的论文，并将每篇文档解析为统一表示形式，在原始阅读顺序中保留散文文本与视觉证据（如图）的原生交替结构。遵循 arXiv 分类体系，SIN-Data 涵盖 12 个以上主要学科和 80 多个子领域，构成广泛的领域知识基础。在 SIN-Data 的基础上，我们开发了一条可扩展的流水线，用于构建 SIN-Bench。每个实例由一个查询、答案以及锚定于原生锚点的证据链构成。作为与真实科学认知相契合的分层评估套件，SIN-Bench 通过四项任务实现渐进式的“发现—验证—综合”工作流程：（1）SIN-Find（证据发现），（2）SIN-Verify（假设验证），（3）SIN-QA（基于依据的推理），以及（4）SIN-Summary（证据锚定综合）。为避免孤立的、只看答案的评估，我们采用“无证据，不评分”原则：只有以可验证证据为依据的高置信度推理才会获得认可。除了答案准确率，我们还从三个维度评估证据质量：逻辑、匹配度和相关性。

实证评估八种广泛使用的多模态大语言模型后发现，证据定位是长篇科学阅读的主要瓶颈。例如，Gemini-3-pro 的平均总体得分最高（0。566），而 GPT-5 的 SIN-QA答案准确率最高（0。767），但在 SIN-Find 和 SIN-QA 的证据对齐总体得分方面表现较差，这表明答案正确并不一定意味着有可追溯的证据支持。此外，对于若干开放权重模型而言，结构化证据输出仍颇具挑战（常导致得分无效）。

总而言之，本文的贡献如下：

- 我们提出“海中之鱼”范式，用于评估基于证据的长上下文多模态科学推理。
- 我们构建了 SIN-Data 和 SIN-Bench，这是一套涵盖发现、验证、质量保证和证据锚定综合的四项任务基准。
- 我们提出“无证据，不评分”原则和多维证据指标，以诊断答案准确率之外的定位质量。

2 相关工作

多模态大语言模型的长上下文理解任务。近期关于多模态大语言模型长上下文理解的研究进展，主要聚焦于探索检索边界。早期范式，如 MM-NIAH （ Wang 等人，2024b ）和 MM-LongCite（Zhou 等人，2025年），采用“干草堆中的针”设置，将合成的文本或图像针插入长上下文中，以测试召回率上限。这些方法有效地量化了上下文容量，但依赖语义孤立的合成内容，无法反映真实认知中的复杂逻辑依赖。为提高真实性，后续基准测试，如 MMLongBench-Doc（Ma 等人，2024年）和 LongDocURL（Deng 等人，2025年），转而采用真实网页和多页报告。然而，这些基准测试主要强调单点信息抽取或页面布局感知，往往忽略了

跨越整篇文档的远距离跨模态推理。为弥补这一不足，我们提出海中之鱼范式。与孤立的检索任务不同，该范式要求构建连贯的跨模态证据链。本研究进一步推出 SIN-Bench，通过模拟完整的科学研究工作流程，评估在高度互联的科学文献中进行深度推理的能力。

评估多模态大语言模型的科学推理。 科学文献包含稠密的多模态内容，因此是评估专家级推理能力的有效领域。一些研究致力于将 PDF 和 LaTeX 文档转换为机器可读格式，为后续研究奠定基础（Wang 等人，2024年a；Zhong 等人，2019年；Xia 等人，2024年）。在此基础上，近期研究转向更高层次的认知。M-DocSum（Yan 等人，2025年）提出了基于参考资料的交错式文本与图像摘要，MMIE（Xia 等人，2025年）则为知识密集型领域设计了信息抽取任务。然而，许多基准测试依赖以预测为导向的指标，例如 ROUGE 或简单问答的准确率，并未评估推理过程的透明度。这种黑箱式评估设置允许答案来自参数记忆，而非文档级理解。受近期深度研究工作的启发（Jiang 等人，2024年；Yang 等人，2025年b），我们提出了一套以证据为中心的指标体系，涵盖匹配度、有效性和逻辑性，并贯彻“无证据，不评分”原则，以便对推理行为进行细粒度诊断。

3 海中之鱼范式

与依赖插入合成噪声的“大海捞针”（Wang 等人，2024b）范式不同，海中之鱼（FITO）范式旨在评估对内在复杂的长篇多模态文档的理解。在这一比喻中，长上下文多模态文档就是典型的“海洋”，具有三个对专家级推理构成挑战的基本属性：（i）**天然性**：信息本就存在且语义相互交织，而非人为注入；（ii）**互联性**：证据横跨不同章节/模态（如文本、图、表格）；（iii）**长程依赖**：逻辑链跨越彼此分离的章节，例如将方法与结果和结论联系起来。

为了评估这一过程，我们不再局限于简单的问答似然最大化，$P(A|D, Q)$，其中A、D和 Q 分别表示答案、文档和查询。相反，FITO 对答案及其支持性证据链（E）的联合概率进行建模：

$$P(A, E|D, Q) = P(E|D, Q) \cdot P(A|E, D, Q). \quad (1)$$

这种表述将评估目标从以结果为导向的预测转向过程导向推理。这意味着，有效的系统必须显式实例化潜变量 E，并在推导出 A 之前验证其充分性。

4 SIN-Bench：基于证据的评估

为了将 FITO 范式付诸实践，我们引入 SIN-Bench。我们选择科学文献作为评估载体，因为其信息密度高、逻辑严谨，适合评估专家级认知能力。具体而言，我们建立了一个半自动化框架，将原始科学语料转化为结构化推理挑战。如图 2 所示，该框架包括构建统一的科学交错（SIN）表示（第 4.1 节）、基于证据的层级任务设计（第 4.2 节）、证据驱动的指标体系（第 4.3 节），以及可扩展的基准构建流程（第 4.4 节）。

4.1 SIN-Data 基础设施

为严格评估长上下文多模态推理，我们建立了 SIN-Data基础设施。这是一条旨在将原始文档统一转换为具备语义感知能力的科学交错（SIN）格式的流水线。我们的首要目标是突破原始格式（如 PDF 布局或原始 LaTeX）的局限，重建一条线性数据流，以保留文本与视觉证据之间基于逻辑而非空间的关联。如图2（a）所示，受近期研究（Wang 等人，2024a）启发，我们的流水线分三个递进阶段处理50k 份来自 arXiv 的原始源数据包[1] 和 PubMed Central（PMC）[2]：**阶段 1：元素解析**。我们将内容提取与呈现层分离。对于 arXiv，我们通过 Engrafo 将源 LaTeX 编译为响应式 HTML，随后进行文本解析。从 DOM 树中提取图像，并通过

[1] https://arxiv.org/ [2] https://www.ncbi.nlm.nih.gov/pmc/

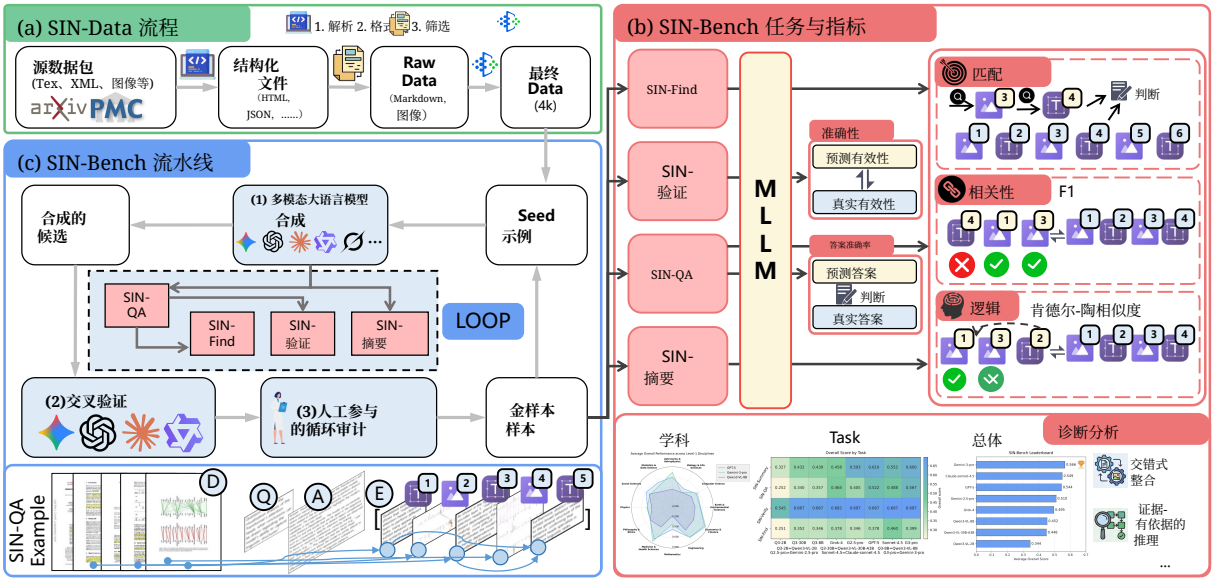


图2：SIN-Bench 的框架。**(a) SIN-Data 基础设施**：我们采用语义优先策略，将原始源数据包解析为统一的科学交错格式。**(c) 构建流程**：基于这些数据，我们采用带有交叉验证和人工审核的迭代式多 MLLM 合成循环，生成高质量样本。**(b) 任务与指标**：该基准测试包含四项层级任务，并按照“无证据，不评分”协议进行评估。我们从匹配度、相关性与逻辑性三个维度评估证据链。

视觉匹配将其重新锚定到文本上下文中。对于 PMC，我们使用 s2orc-doc2json 解析器将 JATS XML 文件解析为结构化 JSON，在剥除装饰性元素的同时保留表格和引文链接。

第 2 阶段：语义优先格式化。结构化数据统一为交错式 Markdown。与依赖版面的 PDF 提取不同，我们提出了引文驱动注入策略，以保留逻辑上的“证据链”。每个视觉组件都会被分配一个唯一 ID $\langle x_k \rangle$，并插入到首次引用它的段落附近（例如，放在“图3表明……”之前）。这确保视觉证据与推理逻辑保持一致。我们还会同时计算质量信号，例如文本与图像交错片段的数量以及词元密度，以供后续处理。

阶段3：质量筛选。为确保多模态密度较高，我们依据质量信号，严格剔除视觉上下文稀疏或引文链接断裂的样本。此外，我们参考 arXiv 官方分类体系对样本进行分类和标注，确保保留能够代表不同领域的分布。

这一流程最终产生 4,000 份高质量文档（D）。这些样本涵盖了多种多样的 12 个顶级学科和超过80个子领域。详情见第 A.1节。

4.2 基于证据的层级任务

为了将 FITO 范式以证据链为驱动的目标转化为可操作的基准实例，SIN-Bench 将专业阅读工作流抽象为四项层级任务：发现 $\rightarrow$ 验证 $\rightarrow$问题回答（QA） $\rightarrow$ 合成。每个实例以单个文档 D 为基础，围绕查询 Q、答案 A和证据链 $E = [e_1, \ldots, e_N]$组织标注，并以交错式锚点 N构成结构，其中奇数索引元素表示视觉锚点，偶数索引元素表示文本片段。这个统一接口确保了四项任务之间的语义对齐，便于进行一致的诊断。

SIN-Find（证据发现）。该任务评估模型在长上下文文档中识别推理路径的能力。给定文档 D 和查询 Q，模型必须构建证据链 E:

$$E = f_{\text{find}}(D, Q). \tag{2}$$

不同于关键词匹配检索， Q 旨在要求复杂推理（例如，因素归纳或方法比较）。因此，模型必须检索语义上充分的跨章节或跨模态信息，并将其组织成逻辑顺序清晰的证据链 E，而非仅仅提取零散片段。

SIN-Verify（假设验证）。此任务旨在考察审计能力，判断所提供的证据是否足以支持特定结论。该任务被表述为一个二分类问题：

$$y = f_{\text{verify}}(D, Q, A, E), \quad y \in \{0, 1\} \tag{3}$$

模型需要判断 E 是否为 A提供了一致且充分的支持。为严格区分，我们通过系统性扰动引入负样本（例如，省略前提或错配条件）。这模拟了真实研究环境中的关键审查流程。

SIN-QA（基于依据的推理）。此任务要求联合生成答案及其来源依据。模型必须同时输出答案 A 和证据链 E：

$$(A, E) = f_{\text{qa}}(D, Q). \tag{4}$$

此处， A 以及 E 中的解释性文本均为全新生成，而非提取所得。这种配置将模型的推理限制在可验证的证据范围内，从而有效区分由证据驱动的理解与基于参数先验产生的幻觉。

SIN-Summary（证据锚定综合）。此任务评估对全文的整体理解和长程整合能力。给定D，模型会生成一份交错式、以证据为锚定的摘要，其中包含多个主张，每个主张均配有相应证据：

$$\hat{S} = \{a_j, E_j\}_{j=1}^{J} = f_{\text{sum}}(D). \tag{5}$$

在我们的交错式响应格式下，整体的“答案”和“证据”彼此重叠：合成可以等价地视为

$$A = \{a_j\}_{j=1}^{J}, \qquad E = \bigcup_{j=1}^{J} E_j, \tag{6}$$

其中每条陈述 a_j 均伴有使其可验证的锚点。通过要求跨部分整合以及为每个主张提供明确证据，该任务将长程依赖和多模态连接性具体化。

4.3 指标：无证据，不得分

为缓解评估偏差，即模型通过错误的推理路径得出正确答案（“歪打正着”现象），SIN-Bench 遵循“无证据，不评分”的哲学。如图 2(b) 所示，我们将评估框架分为两层。对于需要生成证据链的任务（SIN-Find、SIN-QA 和 SIN-Summary），我们采用统一协议来评估证据质量。在此基础上，SIN-QA 还会额外评估答案，而 SIN- Verify 则依据准确率进行评估。

交错式证据链。 我们将真实证据链表示为 $E^\star = [e_1^\star, \ldots, e_N^\star]$，将预测链表示为 $\hat{} E = [\hat{} e_1, \ldots\ldots, \hat{} e_M]$。与统一接口一致，奇数索引处的元素表示视觉锚点，偶数索引处的元素对应文本片段。我们将分组组件定义为：

$$\begin{aligned} K^\star &= \lfloor N/2 \rfloor, & \hat{K} &= \lfloor M/2 \rfloor, \\ (v_i^\star, t_i^\star) &= (e_{2i-1}^\star, e_{2i}^\star), & (\hat{v}_i, \hat{t}_i) &= (\hat{e}_{2i-1}, \hat{e}_{2i}). \end{aligned} \tag{7}$$

在评估过程中，我们将相邻对（v_i， t_i）视为证据的最小单元，并在这一粒度上计算质量指标。

对于 SIN-Find、SIN-QA 和 SIN-Summary 任务，我们通过 MRL（匹配度、相关性与逻辑性）指标评估证据链的质量。我们使用大语言模型作为评估器，以可靠地捕捉长篇证据中的语义等价性，克服传统 n-gram 指标的僵化性。具体而言，M 由大语言模型进行语义判定（例如 Qwen3（Yang 等人，2025a）；模型选择见第 D.3节），而 R 和 L 则依赖确定性的分析公式。

(M) 匹配度。 令 $V^\star = \{v_i^\star\}_{i=1}^{K^\star}$ 表示真实视觉锚点集合。我们将命中集H 定义为预测中视觉锚点与真实锚点匹配的索引子集：$H = \{ i \in [1, \hat{} K] \mid \hat{} v_i \in V^\star\}$。对于每个索引$i \in H$，我们找出相应的真实索引 $j(i)$，使得 $v_{j(i)}^\star = \hat{} v_i$。随后，我们使用大语言模型评估预测文本 $\hat{} t_i$ 与参考文本 $t_{j(i)}^\star$之间的语义一致性，并赋予分数 $s_i \in \{0, 1, 2, 3\}$。我们将此分数归一化为 $\tilde{s}_i = s_i/3$。匹配指标定义如下：

$$\text{Match} = \begin{cases} \frac{1}{|H|} \sum_{i \in H} \tilde{s}_i, & |H| > 0, \\ 0, & |H| = 0. \end{cases} \tag{8}$$

(R) 相关性（F1）。我们将证据单元的正确性表述为一个二元分类问题。正确预测的指示函数为 $\mathbb{I}_i = \mathbf{1}[\hat{} v_i \in V^\star \wedge \tilde{s}_i \geq \tau]$，其中 τ 表示

语义阈值（在我们的实现中设为 $2/3$）。令真正例的数量为$\hat{}TP = \sum_{i=1}^{K} \mathbb{I}_i$，我们计算：

$$\text{Prec} = \frac{TP}{\hat{K}}, \quad \text{Rec} = \frac{TP}{K^\star}, \quad \text{F1} = \frac{2 \cdot \text{Prec} \cdot \text{Rec}}{\text{Prec} + \text{Rec}}. \quad (9)$$

（L）逻辑（肯德尔秩相关相似度）。对于匹配锚点集合，设 $\hat{}\pi$ 和 $\pi^\star$ 分别表示它们在预测结果和真实标注中的相对排列，并且 $K = |H|$ 表示匹配总数。肯德尔秩相关系数（Kendall，1948）的计算如下：

$$\tau = 1 - \frac{2 \cdot \text{inv}(\hat{\pi}, \pi^\star)}{K(K-1)/2} \in [-1, 1], \text{KT-sim} = \frac{\tau + 1}{2} \in [0, 1]. \quad (10)$$

SIN-QA：答案准确率。对于预测的 $\hat{}$答案 A，大语言模型会为其分配语义正确性得分 $g \in \{0, 1, 2, 3\}$。我们对该值进行归一化，以得到答案准确率指标 AnsAcc $= g/3$。**SIN-Verify：准确率**。该任务属于二元分类问题。我们使用 T 样本上的标准准确率评估性能：

$$\text{Score}_{\text{VERIFY}} = \text{Acc} = \frac{1}{T} \sum i = 1^T \mathbf{1}[\hat{y}_i = y_i]. \quad (11)$$

此外，我们通过对各项指标取平均值，计算每个任务的总体得分。

4.4 基准构建流程

基于SIN-Data提供的交错样本，我们设计了一个迭代式人机协作流程。该流程通过迭代合成过程实现可扩展性，同时借助跨模型筛选和人在环审计，确保事实正确性、推理一致性和证据可验证性。如图 2 （c）所示，该流程包括以下关键步骤：

种子示例。我们为每项任务人工创建一小组高质量种子示例。

(1) 多多模态大语言模型合成。给定单个文档 D，我们采用多个多模态大语言模型进行协同合成。我们将 SIN-QA 和 SIN-Summary 指定为核心生成枢纽。模型在单一上下文中联合生成问题 Q、答案A和证据链草稿 E，从而最大限度地减少 SIN-QA 中的语义漂移。随后，我们在同一轮合成中推导出其他任务。我们通过反向执行 SIN-QA 来重新构建 SIN-Find，重点定位支持结论的证据链。对于 SIN-Verify，我们通过系统性扰动 E （例如随机打乱）或创建证据不足但结论表面上看似合理的困难样本，来构造负样本。对于 SIN-Summary，模型通过“边写边引用”提示生成全面的摘要。此步骤产生候选样本池。

(2) 交叉验证。为减轻潜在的合成噪声，我们使用模型集中的三个强大模型进行交叉验证。它们分别从三个维度对候选样本进行 1 到 5 分的评估：问题合理性、答案正确性，以及支持性证据的唯一性和一致性。我们仅在样本获得多数票，且三个维度的评分均为 ≥ 4 时，才予以保留。

(3) 人在回路审计。此流程着重验证证据锚点的精确位置以及支持性关系的严谨性。人类专家还会纠正残留的事实错误和格式偏差，以产出高置信度的金样本。

迭代循环。这些经验证的黄金样本会作为更新后的种子示例反馈到系统中，用于后续循环，直至数据集达到目标规模并保持稳定质量。

利用 4,000 篇高质量科学文档，这一迭代过程生成了约3,200 个原始候选样本。经过多轮循环，我们最终收集到SIN-Bench，包含 490 个样本（查找：159，质量保证：158，摘要：89，验证：84）。更多详情见第A.2节。

5 实验

5.1 实验设置

我们考察了8 种广泛使用的多模态大语言模型，包括5 种专有模型和3 种开放权重模型（详情见第D.1节）。具体而言，5 种专有模型包括：Gemini-3-pro-preview（DeepMind，2025年）、Gemini-2.5-pro（Team，2025年）、GPT-5（OpenAI，2025年）、Grok 4（xAI，2025年）以及 Claude-sonnet- 4.5（Anthropic，2025年）。考虑到开放权重模型架构的多样性，包括稠密架构和混合专家架构，我们选取了3 种广泛使用的模型：启用思考模式的 Qwen3-VL 系列（2B、8B 和 30B-A3B 混合专家版本）（Bai等人，2025b）。

此外，我们在第D.2节详细讨论了所有采用的模型以及选择这些模型的依据。

5.2 主要结果

整体表现。如表1所示，Gemini-3-Pro-Preview 的平均总分最高

模型	SIN-Find				SIN-Verify		SIN-QA					SIN-Summary				Avg.
	L	M	R	O	Acc	O	答案准确率	L	M	R	O	L	M	R	O	总体
专有																
Gemini-3-pro	0.524	0.381	0.294	0.399	**0.697**	**0.697**	0.726	0.569	**0.521**	**0.454**	**0.567**	0.784	**0.546**	0.471	0.600	**0.566**
Claude-sonnet-4.5	0.512	**0.490**	**0.378**	**0.460**	**0.697**	**0.697**	0.708	0.550	0.361	0.333	0.488	0.765	0.493	0.399	0.552	0.549
GPT-5	0.534	0.333	0.268	0.378	0.667	0.667	**0.767**	**0.578**	0.406	0.336	0.522	**0.808**	0.538	**0.483**	**0.610**	0.544
Gemini-2.5-pro	**0.536**	0.300	0.203	0.346	**0.697**	**0.697**	0.684	0.522	0.240	0.197	0.405	0.762	0.547	0.472	0.593	0.510
Grok-4	0.529	0.339	0.266	0.378	0.682	0.682	0.688	0.522	0.371	0.289	0.464	0.699	0.399	0.275	0.458	0.495
开放权重																
Qwen3-VL-8B	0.500	0.293	**0.245**	0.346	**0.667**	**0.667**	**0.476**	**0.514**	**0.341**	**0.234**	**0.357**	**0.683**	**0.359**	**0.276**	**0.439**	**0.452**
Qwen3-VL-30B-A3B	**0.522**	**0.307**	0.226	**0.352**	**0.667**	**0.667**	0.441	0.528	0.321	0.214	0.340	0.676	0.345	0.275	0.432	0.448
Qwen3-VL-2B	0.485	0.160	0.108	0.251	0.546	0.546	0.337	0.510	0.243	0.164	0.252	0.589	0.228	0.165	0.327	0.344

表1：不同多模态大语言模型在 SIN-Bench 上的主要结果。我们报告匹配度（**M**）、相关性（**R**）、逻辑（**L**）、答案准确率（**AnsAcc**）和验证准确率（**Acc**）。**O** 表示任务特定总分，**平均总体得分** 表示所有任务总体得分的算术平均值。最佳表现以**粗体**标示。

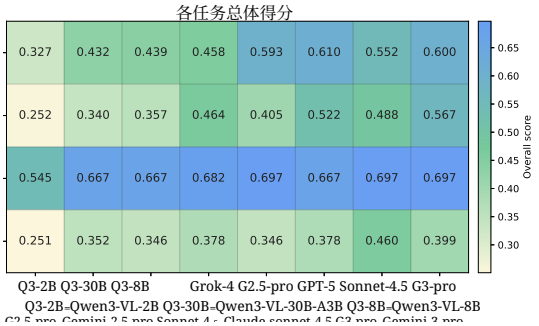


图3：SIN-Bench 中各模型的任务级总体表现热力图。颜色越深，表示得分越高。

（0.566），展现出稳健的证据驱动推理能力。尽管 GPT-5 在 SIN-QA 任务的答案准确率方面领先，但其在 SIN-Find 和 SIN-QA 上的整体表现（总计：0.378/0.522）仍不及 Gemini-3-pro（0.399/0.567）。这一差异表明，Gemini-3-Pro 在多模态上下文中进行定位推理的能力更强。相比之下，GPT-5 的表现表明，它倾向于依赖参数化知识来推断答案，而非严格遵循证据识别流程。更多分析见第 E节。

此外，Qwen3-VL-8B 在几乎所有指标上都优于规模更大的 Qwen3-VL-30B-A3B 变体。这一发现表明，在科学推理领域，推理导向微调的密度比参数总量或混合专家架构更为关键。此外，多个开源模型未能遵守证据格式化约束，导致得分无效，也凸显了在长多模态上下文中生成结构化输出所面临的巨大挑战。

任务级分析。 如图 3所示，我们报告了各模型在不同任务上的热力图。

在**SIN-Find**中，Claude-sonnet-4.5在科学锚点识别方面表现出更高的精确率

设置	**Gemini-3-pro**	**GPT-5**	**Qwen3-VL-8B**
简单负例	1.000	1.000	1.000
困难负例	**0.250**	0.208	0.044

表2：不同负采样设置下 SIN-Verify 的准确率。采用近失证据的困难负例会导致近随机性能。

（总体：0.460），而 Gemini-2.5-pro 在逻辑保持方面表现出色（0.536）。这一模式表明，锚点识别的精确性与证据顺序保持的稳健性之间存在权衡。

在**SIN-Verify**中，标准设置下的准确率仍紧密聚集（Acc=0.667–0。697见表1），但表2 显示，在困难负例下，准确率大幅下降（例如，GPT-5：1.000 →0。208；Qwen3-VL-8B：1.000 → 0.044），这表明模型对近失证据的逻辑辨别能力有限。详细实验见第E.2节。

在 **SIN-QA** 中，Gemini-3-pro 的总体得分最高（0。567），而 GPT-5 的答案准确率最高（0。767），这表明答案正确性与证据质量之间存在差距。这一模式表明，联合解码答案和证据可以提高语义一致性，但不能保证形成蕴含支持链。

在**SIN-Summary**中，GPT-5 的总体得分最高（0.610），且 Logic 和 Relevance 得分最佳。这一结果表明，Logic 指标受益于对科学写作高层次行文脉络的学习（例如，摘要 → 方法 → 结果）。

5.3 讨论

交错输入的重要性。 图。 5表明，保留科学论文的原生交错式结构，相较于分离式布局（图像→文本），可带来显著提升，使 Gemini-3-pro 的 SIN- QA 提高 +0.102，SIN-Summary 提高 +0.129。此外，模态消融表明，“交错式” > “仅文本（图注）” >

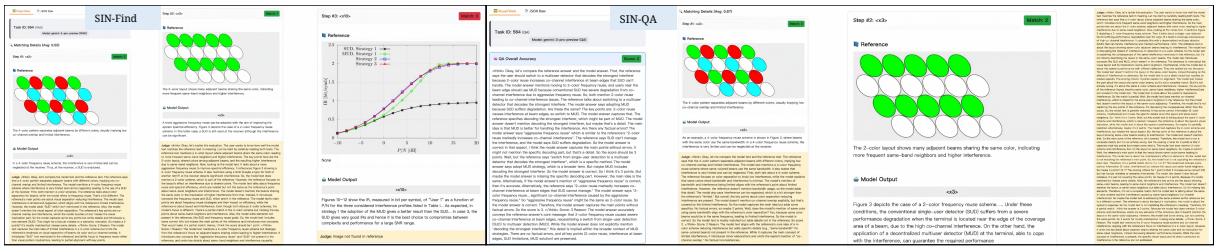


图 4：Gemini-3-pro 在 SIN-Find 和 SIN-QA 任务中的推理失败定性示例。

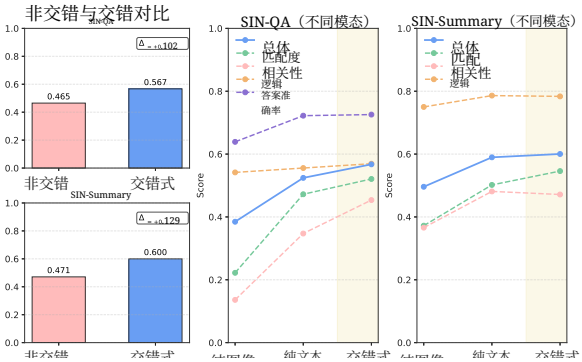


图 5：交错输入和模态编码对 SIN-QA 和 SIN-Summary 的影响（Gemini-3-pro）。

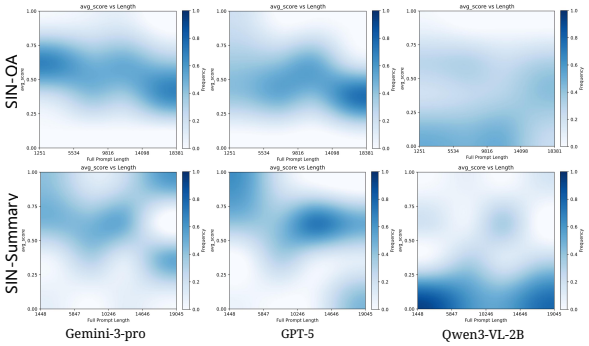


图 6：SIN-QA 和 SIN-Summary 在不同输入词元长度下的分数密度分布。

“仅图像（渲染页面）”，这表明图注保留了粗粒度语义，却丢失了细粒度的实证证据；而当原始视觉内容由相邻文本在局部提供 grounding 时，其效用最大。

“证据链”效应。 我们通过比较 Gemini-3-pro 在 SIN-QA 上启用与未启用证据链生成时的表现，评估要求显式提供证据链是否会影响答案正确性。我们的结果显示，强制输出证据可将性能从 0.694 提升至 0.726。这表明，证据链生成可作为一种轻量级的多模态思维链，促使模型在回答前进行证据锚定，并减少无依据的猜测。

文本长度对性能稳定性的影响。 图6考察了输入上下文超过19k 个文本词元时的推理稳定性（更多分析见第 E.3 节）。Gemini-3-pro 和 GPT-5 均表现出较强的适应能力，不过 GPT-5 在 SIN-QA 上出现了明显的尾部性能下降。相比之下，Qwen3-VL-2B 表现出不稳定性。例如，它的性能急剧崩溃，在 SIN-Summary 中，大量得分集中在零附近。

错误分析。如图4所示，我们的定性评估发现了两种失效模式：（1）信息不足，即模型在逻辑链中遗漏关键前提；（2）虚假推理，其特点是包含无关的“散弹式引用”，从而降低精确率。此外，我们观察到，对于相同的上下文，在执行 SIN-QA 时，模型往往能提供更连贯的推理过程，并减少关键要点的遗漏。这一差异表明，SIN-QA 中的目标答案充当了“语义锚点”，引导模型有效追踪长距离依赖关系。更多分析见第F节。

6 结论

在这项工作中，我们提出了“海中之鱼”范式，并建立了 SIN-Bench，这是一个专为长上下文、交错式科学推理量身打造的综合基准测试。通过“无证据，不评分”机制强制构建明确的证据链，我们的框架揭示了当前最先进多模态大语言模型的一个关键落地差距：GPT-5 等模型往往依赖参数先验，在缺乏真正上下文理解的情况下臆造正确答案；而 Gemini-3-pro 则表现出更强的遵循所提供证据的能力。此外，我们的分析还发现，当前模型存在显著脆弱性，尤其是在应对对抗性的“困难负例”或开放权重架构遵循结构化输出约束时。最终，本研究倡导多模态评估进行根本性转变——超越单纯的回答

准确率，转而严格评估推理过程的可追溯性与逻辑性。

局限性

由于当前多模态架构的输入要求，评估的广度受到限制。尽管本研究纳入了 GPT-5 和 Gemini-3-Pro 等领先的通用模型，但支持长上下文交错输入的模型较为稀缺，因此无法纳入特定领域的专家模型。为缓解这一问题，我们**将所有数据构建与评估代码开源**。这一标准化接口有助于社区整合新兴的多模态大语言模型，从而扩大未来的评估覆盖范围。

此外，为尽量减少解析噪声而设计的严格筛选策略，会在数据利用方面造成权衡，因为存在轻微缺陷的文档也会被无意排除。幸运的是，该流水线配备了高度可配置的筛选规则。这种灵活性允许用户调整阈值和投票机制，从而在纯度与规模之间实现最佳平衡，以检索有价值的文档。

伦理考量

我们严格遵守数据合规原则。SIN-Data 中的所有文档均来自开放获取的知识库，具体包括 arXiv 和 PMC，并且我们遵守知识共享署名 4.0 等相关许可协议。然而，滥用风险值得关注，因为“论文工厂”等实体（Kendall 和 da Silva，2024年）可能会利用这项技术生成伪造论文。尽管“无证据，不评分”原则可作为验证机制，但我们承认，反向滥用可能会制造带有可信证据链的逼真伪造内容。因此，我们敦促社区将研究重点从生成转向利用证据链检测学术欺诈，并鼓励披露技术细节，以促进透明度。

参考文献

Meta AI。2025年。Llama 4 系列：开启原生多模态 AI 创新的新时代。https://ai.meta.com/blog/llama-4-multimodal-intelligence/。发表于2025年4月5日。访问于2025年12月30日。

Anthropic。2025年。Claude Sonnet 4.5 系统卡。Anthropic 系统卡。

Lei Bai、Zhongrui Cai、Yuhang Cao、Maosong Cao、Weihan Cao、Chiyu Chen、Haojiong Chen、Kai Chen、Pengcheng Chen、Ying Chen、Yongkang Chen、Yu Cheng、Pei Chu、Tao Chu、Erfei Cui、Ganqu Cui、Long Cui、Ziyun Cui、Nianchen Deng等共83位作者。2025a。Intern-s1：科学多模态基础模型。*CoRR*，abs/2508.15763。

Shuai Bai、Yuxuan Cai、Ruizhe Chen、Keqin Chen、Xionghui Chen、Zesen Cheng、Lianghao Deng、Wei Ding、Chang Gao、Chunjiang Ge、Wenbin Ge、Zhi- fang Guo、Qidong Huang、Jie Huang、Fei Huang、Binyuan Hui、Shutong Jiang、Zhaohai Li、Mingsheng Li等共65位作者。2025b。Qwen3-VL 技术报告。*arXiv* 预印本 *arXiv:2511.21631*。

Lukas Blecher、Guillem Cucurull、Thomas Scialom 和 Robert Stojnic。2024年。Nougat：面向学术文档的神经光学理解。载于*ICLR*。OpenReview.net。

Zhe Chen、Weiyun Wang、Hao Tian、Shenglong Ye、Zhangwei Gao、Erfei Cui、Wenwen Tong、Kongzhi Hu、Jiapeng Luo、Zheng Ma、Ji Ma、Jiaqi Wang、Xi- aoyi Dong、Hang Yan、Hewei Guo、Conghui He、Bo- tian Shi、Zhenjiang Jin、ChaoXu 等 16 位作者。2024年。我们距离 GPT-4V 还有多远？借助开源套件缩小与商业多模态模型的差距。《中国科学：信息科学》，67(12)。

Pradeep Dasigi、Kyle Lo、Iz Beltagy、Arman Cohan、Noah A. Smith 和 Matt Gardner。2021年。一个以研究论文为依据的信息检索问题与答案数据集。载于*NAACL-HLT*，第4599–4610页。计算语言学协会。

Google DeepMind。2025年。Gemini 3 Pro 模型卡。技术报告，Google。

Chao Deng、Jiale Yuan、Pi Bu、Peijie Wang、Zhong- Zhi Li、Jian Xu、Xiao-Hui Li、Yuan Gao、Jun Song、Bo Zheng 和 Cheng-Lin Liu。2025年。Longdocurl：一项综合性多模态长文档基准测试，整合理解、推理与定位。在*ACL*（*1*），第1135–1159页。计算语言学协会。

Chaoyou Fu、YifanZhang、Shukang Yin、Bo Li、Xinyu Fang、Sirui Zhao、Haodong Duan、Xing Sun、Ziwei Liu、Liang Wang、Caifeng Shan 和 Ran He。2024年。MME-Survey：多模态大语言模型评估综合综述。*CoRR*，abs/2411.15296。

Goeric Huybrechts、Srikanth Ronanki、Sai Muralidhar Jayanthi、Jack FitzGerald 和 Srinivasan Veeravanal- lur。2025年。Document Haystack：一项用于长上下文多模态图像/文档理解的视觉大语言模型基准。*CoRR*，abs/2507.15882。

DongzhiJiang、RenruiZhang、ZiyuGuo、YanminWu、Ji- ayi Lei、Pengshuo Qiu、Pan Lu、Zehui Chen、Guanglu

Song, Peng Gao, Yu Liu, Chunyuan Li 和 Hongsheng Li。2024年。Mmsearch：评估大型模型作为多模态搜索引擎的潜力。*CoRR*, abs/2409.12959。

Graham Kendall 和 Jaime A. Teixeira da Silva。2024年。大型语言模型（如 ChatGPT）在科学出版中的滥用风险：作者身份、掠夺性出版和论文工厂。*Learn. Publ.*，37(1):55–62。

Maurice George Kendall。1948年。秩相关方法。

Hugo Laurençon、Lucile Saulnier、Léo Tronchon、Stas Bekman、AmanpreetSingh、Anton Lozhkov、Thomas Wang、Siddharth Karamcheti、Alexander M. Rush、Douwe Kiela、Matthieu Cord 和 Victor Sanh。2023年。OBELICS：一个开放的网络规模过滤数据集，包含交错排列的图像和文本文档。载于*NeurIPS*。

Yang Liu、Dan Iter、Yichong Xu、Shuohang Wang、Ruochen Xu 和 Chenguang Zhu。2023年。G-eval：使用 GPT-4 实现更好的人类对齐的自然语言生成评估。载于*EMNLP*，第 2511–2522 页。计算语言学协会。

Yubo Ma、Yuhang Zang、Liangyu Chen、Meiqi Chen、Yizhu Jiao、Xinze Li、Xinyuan Lu、Ziyu Liu、Yan Ma、Xiaoyi Dong、Pan Zhang、Liangming Pan、Yu- Gang Jiang、Jiaqi Wang、Yixin Cao 和 Aixin Sun。2024年。MMLONGBENCH-DOC：利用可视化对长上下文文档理解进行基准测试。发表于*NeurIPS*。

Minesh Mathew、Dimosthenis Karatzas 和 C. V. Jawa- har。2021年。DocVQA：一个针对文档图像视觉问答的数据集。发表于*WACV*，第 2199–2208 页。IEEE。

OpenAI。2025年。GPT-5 系统卡。技术报告，OpenAI。

Jonathan Roberts、Kai Han、Neil Houlsby 和 Samuel Albanie。2024年。SciFIBench：针对科学图表解读对大型多模态模型进行基准测试。发表于*NeurIPS*。

Gemini Team。2025年。Gemini 2.5：凭借高级推理、多模态能力、长上下文和新一代智能体能力，拓展前沿。*CoRR*，abs/2507.06261。

Junjie Wang、Yin Zhang、Yatai Ji、Yuxiang Zhang、Chun- yang Jiang、Yubo Wang、Kang Zhu、Zekun Wang、Tiezhen Wang、Wenhao Huang、Jie Fu、Bei Chen、Qunshu Lin、Minghao Liu、Ge Zhang 和 Wenhu Chen。2024a。PIN：一个面向成对及交错多模态文档的知识密集型数据集。*CoRR*，abs/2406.13923。

Weiyun Wang、Shuibo Zhang、Yiming Ren、Yuchen Duan、Tiantong Li、Shuo Liu、Mengkang Hu、Zhe Chen、Kaipeng Zhang、Lewei Lu、Xizhou Zhu、Ping Luo、Yu Qiao、Jifeng Dai、Wenqi Shao 和 Wenhai Wang。2024b。多模态大海捞针。发表于*NeurIPS*。

xAI。2025年。Grok-4 模型卡。模型卡，xAI .

Peng Xia、Siwei Han、Shi Qiu、Yiyang Zhou、Zhaoyang Wang、Wenhao Zheng、Zhaorun Chen、Chenhang Cui、Mingyu Ding、Linjie Li、Lijuan Wang 和 Huaxiu Yao。2025年。MMIE：面向大型视觉语言模型的大规模多模态交错理解基准。在*ICLR*。OpenReview.net。

Renqiu Xia、Song Mao、Xiangchao Yan、Hongbin Zhou、Bo Zhang、Haoyang Peng、Jiahao Pi、Daocheng Fu、Wenjie Wu、Hancheng Ye、Shiyang Feng、Bin Wang、Chao Xu、Conghui He、Pinlong Cai、Min Dou、Botian Shi、Sheng Zhou、Yongwei Wang及其他4位作者。2024年。《Docgenome：用于训练和测试多模态大语言模型的开放式大规模科学文档基准测试》。*CoRR*，abs/2406.11633。

Haolong Yan、Kaijun Tan、Yeqing Shen、Xin Huang、Zheng Ge、Xiangyu Zhang、Si Li和Daxin Jiang。2025年。《M-docsum：大型视觉语言模型是否真正理解文档摘要中的交错图文？》*CoRR*，abs/2503.21839。

AnYang、Anfeng Li、Baosong Yang、BeichenZhang、Binyuan Hui、Bo Zheng、Bowen Yu、Chang Gao、Chengen Huang、Chenxu Lv、Chujie Zheng、Day- iheng Liu、Fan Zhou、Fei Huang、Feng Hu、Hao Ge、Haoran Wei、Huan Lin、Jialong Tang及其他40位作者。2025a。《Qwen3技术报告》。*CoRR*，abs/2505.09388。

Zhaorui Yang、Bo Pan、Han Wang、YiyaoWang、Xingyu Liu、Minfeng Zhu、Bo Zhang 和 Wei Chen。2025年b。多模态深度研究者：使用智能体框架从头生成文本与图表交错报告。*CoRR*，abs/2506.02454。

Lianmin Zheng、Wei-Lin Chiang、Ying Sheng、Siyuan Zhuang、Zhanghao Wu、Yonghao Zhuang、Zi Lin、Zhuohan Li、Dacheng Li、Eric P. Xing、Hao Zhang、Joseph E. Gonzalez 和 Ion Stoica。2023年。使用 mt-bench 和聊天机器人竞技场评判大语言模型裁判。发表于*NeurIPS*。

Xu Zhong、Jianbin Tang 和 Antonio Jimeno-Yepes。2019年。Publaynet：迄今规模最大、用于文档版面分析的数据集。发表于*ICDAR*，第 1015–1022 页。IEEE。

Keyan Zhou、Zecheng Tang、Lingfeng Ming、Guanghao Zhou、Qiguang Chen、Dan Qiao、Zheming Yang、Libo Qin、Minghui Qiu、Juntao Li 和 Min Zhang。2025年。Mmlongcite：评估长上下文视觉语言模型保真度的基准测试。*CoRR*，abs/2510.13276。

附录

A 基准创建详情

A.1 SIN-Data 构建流程

设计理念。 我们有策略地将目标聚焦于科学交错（SIN）格式，以应对当前多模态训练和评估中高保真推理数据的稀缺问题。尽管 OBELICS 等大规模数据集（Laurençon 等人，2023年）提供了海量的交错式序列，但这些序列主要源自通用网络资源，其中文本与图像之间的语义对齐往往较为松散或仅属偶然。相比之下，科学文献展现出严格的协同依赖关系：文本提供分析性叙述，而图则提供实证证据。这一领域在内容密度方面具有独特优势，因为科学写作本质上知识密集，信噪比很高。它要求使用精确的术语和逻辑，而非泛泛的描述。

此外，交错式结构对于培养长上下文推理能力至关重要。在科学语境中，解读单幅图表往往需要综合散见于多页稠密散文文本中的信息。我们通过在叙述中以逻辑方式嵌入视觉内容，保留这种非线性的“证据链”。这种设计促使模型在较长上下文中进行多跳推理，而标准的以图注为中心的数据集中往往缺少这种能力。

为了发挥这一潜力，SIN-Data 流程经过专门设计，旨在弥合不同学科之间的结构异质性，尤其是物理学和计算机科学中占主导地位的基于 LaTeX 的源数据包，与生物医学中常见的 XML 结构化档案之间的差异。我们的流水线将这些不同来源统一为标准化格式，同时提供精细的可控性。在同质化过程中计算词元长度和视觉密度等丰富的质量信号后，我们将数据集从原始集合转变为结构化资源。这使我们能够针对不同程度的多模态复杂度，对模型性能进行精确分层和分析。

我们的灵感来自 PIN 数据集（Wang 等人，2024a），我们据此重新设计了解析流程，并纳入质量信号，以构建我们的SIN-Data数据集。如图2 (a) 所示，我们的流程分为三个递进阶段：

A.1.1 阶段 1：元素解析

初始阶段侧重于将内容与呈现形式分离，以生成中间结构化文件。

• arXiv 数据流：我们使用 Engrafo[3] 将 LaTeX 源代码包转换为自适应 HTML，随后通过 NOUGAT（Blecher 等人，2024年）解析以提取文本。视觉匹配算法从 DOM 树中恢复图像引用，并将其重新锚定到对应的文本上下文。

•PMC 数据流：我们利用定制的 s2orc-doc2json[4] 解析器处理 JATS XML 文件。该模块能够稳健地提取核心学术元素——舍弃样式标记，同时保留参考文献链接和表格——并输出结构化 JSON 文件。

A.1.2 阶段 2：语义优先格式化与信号提取

我们将中间文件转换为统一交错 Markdown 格式。一项核心创新是我们的引文驱动注入策略：在图像首次被引用的段落前立即插入唯一的图像占位符 $\langle x_k \rangle$，从而保留逻辑上的“证据链”，而非空间布局。此外，在这一线性化过程中，我们会为每篇文档计算一组细粒度的质量信号。

• 篇幅与粒度：我们记录总计_词元（通过 Llama-4 分词器计算（AI atMeta，2025年））以及平均值_片段_长度（每个文本块的平均词元数），以衡量上下文长度和文本密度。

• 视觉密度：我们计算图像_数量和图像_比例，以量化视觉信息所占的比例。

• 互联性：我们跟踪交错_片段，以表示文本与图像交集的频率，并将其作为多模态推理链复杂度的替代指标。

• 布局与分辨率：我们还提取了特定于格式的特征，例如列布局（单栏或双栏）和平均图像分辨率，以评估视觉质量。这些信号以元数据形式嵌入，作为后续筛选的定量依据。

[3] https://github.com/arxiv-vanity/engrafo[4] https://github.com/allenai/s2orc-doc2json

A.1.3 阶段3：质量筛选与分类体系对齐

利用第 2 阶段计算出的质量信号，我们依据高密度多模态互联原则整理数据集。我们会筛除视觉上下文稀疏、引用链断裂或长度极端的样本（保留32k–1M 词元）。为确保领域覆盖范围广泛，我们参考 arXiv 分类体系对所有样本进行分类，并使用 Qwen3-VL-2B（Bai等人，2025年b）（详情见第B.4节）。我们采用跨类别的分层抽样策略，以防止领域坍缩，确保最终数据集在不同科学领域中保留具有代表性的分布。

从最初收集的 50，000 份源资料包中，我们通过严格的筛选流程，筛选出 4，000 份经过整理的高质量文档。最终语料库涵盖 $10+$ 个顶级学科和超过 $80+$ 个子领域，为科学推理提供了多样且结构可控的测试平台。

A.2 SIN-Bench 构建流程

基于SIN-Data中的交错式文档表示，我们建立了一个用于构建SIN-Bench的人机协作迭代流程。该框架遵循尽量减少人工干预、同时严格保证样本有效性的原则。如图2 (c) 所示，为实现可扩展性和可靠性，我们采用结构化合成、跨模型一致性检查以及人机协作协议。这些措施确保事实准确、逻辑自洽，并使证据可验证。如图所示，工作流程包含四个依次进行的阶段和一个自我改进循环：种子示例收集 $\rightarrow$ (1) 多模型合成 $\rightarrow$ (2) 交叉验证 $\rightarrow$ (3) 人机协作审查 $\rightarrow$ 黄金样本到种子样本迭代。

输入准备：裁剪与对齐。鉴于科学文献通常篇幅很长，我们会进行策略性裁剪，生成符合模型上下文长度限制的输入片段，同时严格保留 SIN 格式的逻辑耦合关系。我们优先保留摘要、引言、方法、实验和结论等核心章节，并保留图或表首次被引文引用处附近的文本片段，以及相应的视觉锚点 $\langle x_k \rangle$。这一过程确保关键论断及其相关证据锚点即使在输入被压缩后仍可定位和验证。

种子示例。 我们为四项任务分别人工整理了一小组高质量种子示例，以统一输出格式和证据链接口。

(1) 多 MLLM 协作合成。 给定单个文档 D，我们利用多种多模态大语言模型协作合成候选样本，包括 Gemini-3-pro、GPT-5、Qwen3-VL、Grok-4 和 Claude-4.5。我们将 SIN-QA 和 SIN-Summary 指定为核心生成枢纽。通过要求模型在统一上下文窗口中共同生成内容（答案或摘要）和证据链E，我们减少了传统流水线中常见的语义漂移和自洽性不足问题。随后，我们在同一合成会话中同步进行任务派生，以从结构上保证跨任务一致性：

- **SIN-QA**：模型针对特定推理查询 Q生成答案 A和证据链 E 。这一过程确保生成的推理依据与源底层材料之间严格语义对齐。

- **SIN-Find**：我们将核心任务的有效输出重新表述为证据定位挑战。通过固定查询 Q 或摘要中的特定主张，我们要求重建支持性证据链。

- **SIN-Verify**：我们对有效证据链应用负采样，构造具有判别性的实例。我们引入“不充分证据”（将有效主张与无关证据配对）和“扰动证据”（打乱有效证据链的顺序或对齐关系），以测试模型审查逻辑缺口的能力。

- **SIN-Summary**：采用“边引用边写作”策略，模型生成整体摘要，其中每个关键陈述都明确链接到 D 中的可验证证据锚点，从而将基于依据的长上下文合成付诸实践。

此阶段会生成一个候选样本池，涵盖全部四项任务。它在各样本之间保持结构耦合，从而便于后续步骤进行统一质量控制。

交叉验证审查员提示：SIN-Summary

角色：
你是一位严谨的学术审稿人，负责审查根据研究论文生成的摘要的**证据质量**。你的判断必须**以证据为中心**，并且**严格以**所提供的摘录为依据。

输入（你将收到）：
1) 一段研究论文的 Markdown 摘录（可能包含与图像/图相对应的 <图> 占位符）。2) 一组摘要证据条目。每个条目包含 (i) 一个图像锚点标记（例如 x1、x2、.....）和 (ii) 一段描述要点的文本片段。

任务描述：
评估所提供的证据条目是否忠实地支持所声称的摘要要点，并明确验证**锚点与内容是否对齐**（即每个 xk 是否正确对应摘录中的相关视觉内容及其周边文本上下文）。

评分标准（按标准细分，1–5）：
证据质量（仅限整数分数）
- **5**：完全忠实且支持充分；锚点匹配准确；涵盖关键方面（方法、结果、 ts, 结论）；组织结构连贯。
- **4**：总体忠实，只有少量遗漏或锚点论证略显薄弱；没有关键性不匹配。- **3**：部分内容得到支持；覆盖范围存在明显缺口，或锚点相关性不清晰；部分证据关联较弱。- **2**：支持力度较弱；多项内容无关、不匹配，或未能在摘录中找到充分依据。- **1**：缺乏依据或包含幻觉；锚点与所述内容不匹配；证据大多无效。

评估清单（证据优先）：
- 文本片段是否准确反映了摘录中的关键要点？- 引用的图像锚点（x1、x2、.....）是否与对应的文本片段相关？- 证据条目是否涵盖了重要方面（方法、结果、结论），而非表面细节？- 证据列表是否组织得当、顺序合理（例如，方法 →实验 →结论）？- 判断是否**仅**依据所提供的摘录（不使用外部知识）？

输出格式（严格遵循）：证据质量评分： <整数 1-5>理由： <一段论证文字>

重要限制：
- 评分必须单独占一行，并以 Evidence Quality Score: 开头。- 评分必须是 1 到 5 的整数（不得使用小数）。- 除规定的两行外，不得输出任何其他字段、项目符号列表或额外行。- 不得使用中文字符或花括号 { }。- 如果某个条目引用的锚点未得到摘录内容的支持，必须相应扣分。

输出示例：
Evidence Quality Score: 4
原因：证据条目大多有摘录内容作为依据：条目 1 正确引用 x1 来描述所提出的架构，且与上下文相符；条目 2 引用 x2 来说明主要实验对比，报告的增益也准确。不过，条目 3 的锚点与主张之间的关联交代不够充分，因为所引用的 x3 与所述消融结论仅有松散关联。尽管整体组织连贯，这仍降低了总体忠实度。

图7：用于审核 SIN-Summary 证据链的交叉验证提示。该提示将证据优先评估标准转化为离散的 1–5 分档，并强制采用严格且可由机器解析的两行输出格式，从而能够大规模筛选匹配度不佳或依据薄弱的图像—文本证据。

(2) 交叉验证。 为系统地筛除看似合理却无法验证的实例，我们采用三模型评审系统，成员从 Gemini-3-pro、GPT-5、Claude- 4.5 和 Qwen3-VL-Plus 中动态抽取。每个模型独立审核候选样本，并在三个维度上分别进行 1 至 5 分的整数评分：问题有效性 （评估清晰度、非平凡性和文档支持情况）、答案正确性 （确保事实一致性，且不存在严重幻觉或遗漏），以及证据一致性（验证相关性、充分性和逻辑顺序）。为确保高质量，我们执行严格的准入标准：只有获得评审系统多数成员批准，且在每个评估维度上均达到至少 4 分的样本，才会被保留。

交叉验证的核心目标并不是

进行冗余生成，而是利用基于裁决、可解析且以证据为中心的提示词，将主观审查转化为可扩展的筛选机制。我们在所有验证提示中统一采用以下设计原则：

• **证据优先**：我们明确要求审查围绕证据锚点及其支撑关系展开，尽量减少对语言流畅度或表面合理性的偏见。

• **标准化评分细则**：我们将评估分解为固定维度和离散等级（1–5），以减少不同审查模型之间的校准偏差。

• **严格输入/输出模式**：我们强制采用固定的输出字段和整数评分，以便进行自动解析和基于阈值的筛选，避免自由格式回复带来的噪声。

• **锚定验证**：我们明确验证视觉标识符（例如$\langle x_k \rangle$）与文本片段之间的映射，以检测图文不匹配和无关引文。

•**防幻觉**：我们要求判断严格依据所提供的文档片段和锚点中的支持证据，防止根据外部参数知识补全内容。

以图7为例，SIN-Summary 的验证提示要求审查员关注“证据质量”，检查摘要条目是否准确涵盖论文要点、引文锚点是否仍然相关，以及内容组织是否符合逻辑顺序。

(3) 人在回路审计。 尽管通过了交叉验证，候选样本仍可能存在细粒度的依据偏差，例如锚点虽可定位，却缺乏严谨的逻辑支持，或遗漏关键前提。为解决这一问题，我们组建了一支由24名研究生和博士生组成的审查团队。每个样本均由2–3名审阅者独立验证，任何分歧则由资深研究人员裁定。人工审查严格遵循以下标准：

• **定位锚定性**：每个$\langle x_k \rangle$ 和文本片段都必须能够在 SIN 文档中精确定位，以确保引文明确无歧义。

• **支持的严谨性**：证据链必须充分且必要，才能支持结论。对于 SIN-Verify 中的负样本，审阅者需确认其确实具有信息不足或前后不一致的属性。

• **事实一致性**：答案或摘要不得引入文档之外的信息。数值、比较性结论和条件设定必须与源文本严格一致。

• **结构合规性**：输出必须符合任务格式和交错式证据链的规范，确保证据顺序与推理过程的逻辑依赖关系一致。

• **可区分性**：问题应避免过于简单或无法判定。证据链必须具有诊断价值，能够有效区分证据驱动推理与侥幸猜测。

这一阶段会生成高置信度的金样本，并为更新种子示例提供可靠信号。

（4）自举迭代（金样本 → 种子）。金样本会回流至系统，作为下一轮合成与推导的更新种子示例。这形成了一个持续运行的反馈闭环，直至基准测试达到规模目标并确保质量稳定。

规模与构成。基于由 4，000 份高质量科学文档构成的语料库，多轮合成流程生成约 3，200 个原始候选样本。经过三模型交叉验证和人在环审查后，SIN-Bench 最终包含 490 个金样本：159 个用于 SIN-Find，158 个用于 SIN-QA，89 个用于 SIN-Summary，84 个用于 SIN-Verify。

此外，目前发布的金样本数量受到人工审核成本的限制，因此高质量样本的数量目前有限。不过，得益于所提出的半自动化协作流程，该基准测试具备内在的可扩展性。只要有充足的计算和人力资源，该流程在样本数量方面就不会遇到瓶颈。

Task	平均长度	最大长度	最小长度
SIN-Find	108.89	186	60
SIN-QA	122.77	203	78
SIN-Verify	243.49	322	160

表3：不同任务类型下输入上下文的词元长度统计。

范围	值
词元（文本 / 图像 / 总计）	15k / 3k / 18k
图像比例	15%
文本比例	85%
每个实例的平均图像数	6.6

表4：基准测试输入中每个样本的总体模态构成。

Task	粗体平均值	斜体平均值	标题平均值
SIN-Find	70.52	122.46	14.64
SIN-QA	70.83	118.17	15.03
SIN-Summary	37.46	155.74	14.91
SIN-Verify	90.35	155.65	15.94
总体	**68.01**	**132.81**	**15.03**

表5：各任务的标记统计（粗体/斜体/标题词元），反映了输入中的结构化格式线索。

B 关于“无证据，不评分”原则

“无证据，不得分”是一项以证据为中心的评估原则：如果模型的输出无法依据可验证的文档锚点进行佐证，就不应仅凭猜测答案获得较高的任务得分。因此，对于要求提供证据的任务，证据质量被视为最终得分的核心组成部分，而非可有可无的解释。

至关重要的是，该原则针对的是可验证性，而非引用表层形式。实际上，只要锚点可以恢复并与论文中的索引视觉标记匹配，我们就不会惩罚细微的格式变化（例如，“Fig. 2”、“Figure 2”或“x2”）。只有当预测未提供任何可恢复的锚点（例如，缺失或无效的标识符）时，证据得分才为零。

我们还阐明了该原则如何与单独的答案评分在逻辑上保持一致。我们始终将答案准确率（AnsAcc）作为对预测答案的独立评估来计算，但仅有答案准确率不足以在没有证据的情况下获得较高的整体任务得分。当没有匹配到锚点时，证据指标按定义降为零（例如，匹配度 $= 0$、F1 $= 0$；若 KT 相似度因匹配不足而未定义，我们则将其设为 $= 0$）。这明确限制了仅凭答案正确所能带来的贡献，并确保要取得有竞争力的表现，就必须提供有依据的证据；这正是“无证据，不评分”原则所期望的效果。

最后，现有指标已涵盖长篇科学文献定位中常见的实际边界情况。无关锚点的过度生成会体现为相关性精确率降低；必要证据覆盖不足会体现为召回率降低；而看似合理但经过挑选或顺序错误的证据，则会体现为逻辑指标下降。这些行为均由正文中定义的指标体系处理，无需引入额外指标。

B.1 总体统计

我们报告基准测试输入的描述性统计数据，重点关注 (i) 实例长度、(ii) 文本-图像构成，以及 (iii) 结构标记密度。

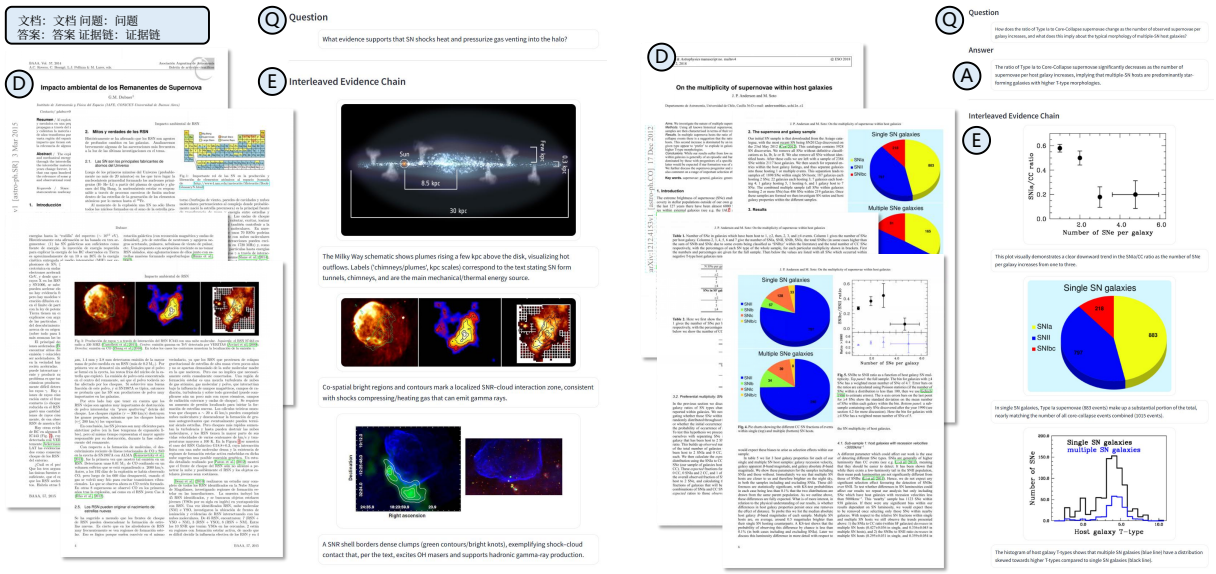
表3 总结了按任务类型划分的词元长度统计。总体而言，SIN-Verify 实例的长度明显超过 SIN-Find 和 SIN-QA（平均约为其2 倍），这表明其上下文和推理负荷更高，而 SIN-Find 和 SIN-QA 的长度范围则相近。

表 4描述了整体模态构成。平均而言，图像相关词元占总词元数的15% ，每个实例包含6.6张图像，这表明任务需要一定程度的视觉定位，同时仍以文本驱动为主。

最后，表 5 报告了结构标记统计数据（粗体/斜体/标题词元），这些数据反映了输入中格式线索和章节划分信号的密度。这些线索会影响模型解析文档结构的方式，并可能影响不同任务类型的难度和计算成本。

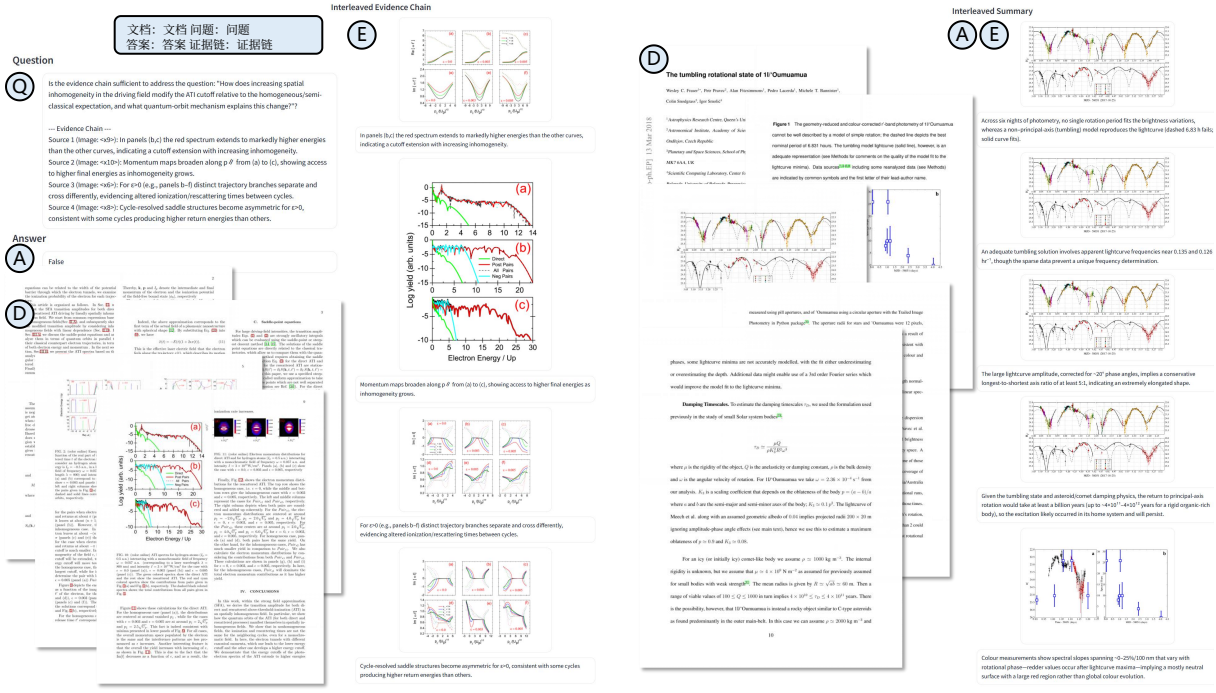
B.2 SIN-Bench 样例

如图8 和图9所示，我们在SIN-Bench中为每项任务展示一个“金样本”实例。由于科学文档篇幅较长，呈现完整实例会浪费空间，也会削弱与任务相关的信号。因此，我们仅展示覆盖文档、问题、标注的证据区域以及标准输出（答案、验证决策或摘要条目）的裁剪后的截图。基准测试本身基于完整的交错源文件（包含图像的 Markdown



SIN-Find　　SIN-QA

图8：精　　选自 SIN-Bench 的黄金示例：SIN-Find 和 SIN-QA。我们为每项任务展示一个精选实例，并用明确的
标记标示**文**档、**问**题、**答**案以及黄金**证**据链。



SIN-Verify　　SIN-Summary

图9：SIN-Bench 中 SIN-Verify 和 SIN-Summary 的金样本示例。我们为每项任务展示一个精选实例，并明确标注**文档**、**问题**、**答案**和金样本**证据链**。

资源），这些文件随附稳定标识符，可供精确重建。

B.3 许可

基准测试数据集和源代码以MIT许可证发布。许可证全文请见
https://opensource.org/licenses/MIT。

B.4 学科分类列表

为全面评估跨科学学科的多模态推理，我们参考arXiv学科分类构建了三级层次分类体系。

我们的流水线将大模型的能力与专家人工验证相结合。我们首先使用 Qwen3-VL-2B 对收集到的样本进行初步分类，

随后，领域专家审核结果，纠正错误分类并完善子领域定义。

最终形成的分类体系包含 12 个一级主要学科，分为 35 个二级领域和 84 个三级子领域。层级结构如下：

- **天文学与天体物理学**：
  – 河外天文学与宇宙学：活动星系核与类星体；星系形成与演化。– 星系演化与结构：星系结构与演化。– 行星与空间科学：行星地质学与探测。–行星系统与形成：行星系统、原行星盘。

- **生物学与生命科学**：
  – 神经科学：神经影像学。
- **计算机科学**:– 算法与理论：组合算法；图与组合算法；优化；量子算法与计算理论。– 通信与网络：信息论与编码；信道建模。– 计算机视觉：二维与三维重建和建模；图像处理与复原；视频理解与压缩；视觉。– 电气与计算机工程：嵌入式与硬件系统。– 人机交互：虚拟与增强现实及触觉技术。– 机器学习与人工智能：主动学习；聚类与无监督学习；深度学习；深度强化学习；差分隐私与匿名化；可解释与公平人工智能；联邦与持续学习；生成模型；图神经网络；多模态学习与融合；非线性时间序列；目标检测；粒子滤波与数据同化；隐私；量子算法与计算理论；二阶方法；次模优化。– 自然语言处理：计算语言学；信息检索与质量保证系统。– 机器人学：自主与移动机器人；机器人控制学习。–安全与隐私：对抗机器学习；区块链与分布式账本安全；差分隐私与匿名化

化。– 软件工程与编程：面向软件工程的人工智能；程序综合与验证。– 系统与基础设施：边缘与云计算及无服务器计算。

- **地球与环境科学**：
  – 气候与大气科学：大气化学与空气质量；气候建模与预测。– 地球物理学与遥感：地理空间智能与 GIS；卫星与航空成像。– 海洋学与水文学：水资源与水文学。– 行星与空间科学：行星地质学与探索；空间仪器与任务；空间天气与近地环境。

- **经济学与金融学**：–宏观经济学与国际贸易：主权债务与违约风险。

- **工程学**：– 机械与机器人工程：机器人与自动化。

- **数学**：–纯数学：分析学与泛函分析。

- **医学与健康科学**：– 神经病学与精神病学：计算精神病学。

- **哲学与伦理学**：– 人工智能与技术伦理：量子物理学哲学。

- **物理学**：
  – 天体物理学与宇宙学：天文仪器；大气化学与空气质量；宇宙学与星系形成；系外行星与行星科学；引力透镜与宇宙学；引力透镜与微引力透镜；中子星与脉冲星；行星地质学与探测；卫星与航空成像；空间仪器与任务；超新星遗迹与宇宙射线；超新星与恒星演化。–凝聚态物质与材料：计算材料物理学；凝聚态物理学；材料科学与表征；自旋玻璃与磁性材料；超导与磁性。– 光学、流体与动力学：激光物理与光子学；光子晶体与带隙物理；等离子体与聚变物理；量子

光学与原子物理学。– 光学与光子学：非线性光学与光学处理；光学成像与传感。– 量子与粒子物理学：引力波与宇宙学；高；中微子天文学与天体物理学；量子计算与信息；量子光学与原子物理学。

- **社会科学**：
  – 传播、媒体与语言学：计算社会媒体分析。– 经济学与金融学：计量经济学与应用微观经济学。– 教育与学习科学：教育技术与教育人工智能。– 社会学与社会政策：社会网络与计算社会科学。

- **统计学与数据科学**：
  – 数据科学应用：健康与社会数据科学。– 统计理论与方法：非参数统计与贝叶斯统计。

这种结构化分布确保 SIN-Bench 覆盖广泛的知识领域。

C 与现有基准测试的比较

表 6将 SIN-Bench 置于具有代表性的长文档和科学多模态基准测试之中。以往的基准测试在长上下文压力测试和以文档为中心的任务覆盖方面取得了显著进展，包括长 PDF 理解套件（如 MMLongBench-Doc）以及将稀疏目标插入长上下文中的针式检索测试（如 Document Haystack）。科学领域的资源则通过结构化文档任务（DocGenome（测试集））或聚焦图表的评估（SciFIBench）进一步拓展了覆盖范围。尽管取得了这些进展，大多数现有基准测试仍主要依赖答案层面的评分，并未明确要求模型生成可审计的跨模态证据链，因此难以区分基于文档依据的推理与看似合理却缺乏依据的生成。

相比之下，SIN-Bench 针对以交错式图像和文本格式呈现的原生科学文档，评估基于证据的科学理解能力，并组织了四项循序渐进的任务（查找/验证/问答/摘要），对应一个发现→验证→合成工作流程。除了答案正确性，我们还明确评估证据——质量（例如匹配度/相关性/逻辑性），从而能够细致诊断失败的原因（是缺少必要前提，还是引文不可靠），而非将其简化为仅依据答案的评估信号。尽管 SIN-Bench 评估的文档数量少于某些大规模资源，但它从最初收集的 5 万篇论文中提炼而来，经过多阶段的清理和筛选流程，旨在确保每一步都保留高质量的科学文档。重要的是，这一流程具有可扩展性：人工参与程度较低，主要限于验证（是非/序数判断），无需大量手动编写问题和参考答案，同时仍能生成可靠且以文档为依据的评估实例。

文档 实验设置详情

文档.1 基线模型

我们报告了所评估的所有专有模型的版本标识符，以支持结果复现。我们对所有模型（包括开放权重模型）均仅进行一次评估，温度设为 0。所评估的专有模型及其版本标识符如下：

- Gemini-3-pro: gemini-3-pro-preview-11-2025
- Gemini-2.5-pro：gemini-2.5-pro-thinking
- Claude-sonnet-4.5: claude-sonnet-4-5-20250929-thinking
- GPT-5：gpt-5
- Grok-4：grok-4-0709

文档.2 基准测试和实验中使用的模型

我们在 SIN-Bench 中评估了专有模型和开源模型的组合。在分析实验中，计算预算和迭代成本限制了受控消融实验的数量，因此实验设计优先选择一个具有代表性的强大模型作为主要测试平台。Gemini-3-pro 作为这一主要模型，因为在跨任务平均计算后，它在 SIN-Bench 中取得了最佳的总体性能，使其成为诊断分析的代表性参考点。

在图中的模态研究5 （第5.3节）中，我们通过构建

基准测试	文档类型	规模	平均长度	领域覆盖范围	主要任务	Ev.	证据链得分
MMLongBench-Doc（Ma 等，2024年）	科学PDF文档（多领域）	130–135份文档（基准测试）	~21千词元	7 个领域（例如：学术/法律/技术）	长文档理解；数值推理；跨元素搜索	No	No
DocGenome（测试集）（Xia 等人，2024年）	科学论文（结构化文档）	–	–	8个类别 +153 子类	7项文档任务（例如，分类/定位/质量保证/版面检测，.....）	No	No
Document Haystack（Huybrechts 等人，2025年）	长篇PDF文档（通用）	400 份文档	5–200 页	通用	大海捞针检索（长文档）	No	No
SciFIBench（Roberts 等人，2024年）	科学图表（图表）	–	–	8 种图类型（以计算机科学为中心）	图表解读	No	No
SIN-Bench（我们的工作）	科学文档（交错式）	231份文档 / 490个实例	~18k 个词元	10+ L1领域；30+ L2； 80+L3	4 项任务：查找/验证/质量保证/摘要（发现→合成）	**Yes**	**Yes**

表 6：与具有代表性的长文档和科学多模态基准测试的比较。“证据”表示该基准测试是否要求模型输出中包含明确的证据链，“证据评分”表示是否对证据质量进行明确评估（而不仅仅是对答案评分）。

一种“纯文本”设置，来分离原始视觉输入的影响。该设置将原始交错序列中的每个图像位置替换为相应图的自然语言描述。这些描述由 Qwen3-VL-8B-Instruct 生成，在移除直接视觉信号的同时保留文档结构。

对于需要语义判断的评估环节，我们采用基于大语言模型的评估器，而非 n 元语法重叠指标。这种设计降低了对改述和表面形式变化的敏感度；尽管语义等价，n 元语法指标往往仍会对此予以惩罚。它也适用于存在多种有效答案表述的场景，因为单一参考文本可能无法充分界定正确性，从而导致系统性假阴性。此外，大语言模型判断器支持由评分标准指导的基于准则的评分，与任务定义直接一致。先前研究表明，基于评分标准或表单的大语言模型评估能够提升与人类判断的一致性，并为开放式输出提供可扩展的评估（Liu et al., 2023；Zhengetal., 2023）。具体而言，匹配度和答案正确性（答案准确率）使用 Qwen3-8B 作为自动判断器。此外，在第 D.3 节中，我们将该判断器与人类专家判断进行基准测试，发现两者高度一致，这支持了基于大语言模型的评估器用于大规模评估时的可靠性。

D.3 判断指标

为确保自动评估在匹配度和问题准确率指标上的可靠性，我们采用 Qwen3-8B 作为判断。我们通过一项严格的一致性研究验证了这一选择，该研究涵盖每项任务的所有实例。对于这些样本，我们采用双重评分方案，将自动评判器与人类专家进行比较。

Task	指标	皮尔逊（r）	斯皮尔曼（ρ）
SIN-Find	匹配度（M）	0.824	0.795
SIN-QA	匹配（M） 答案准确率	0.812 0.895	0.788 0.862
SIN-Summary	匹配（M）	0.768	0.741
-	**平均值**	**0.825**	**0.797**

表7：不同任务和指标下，人类专家评分与自动评分（由 Qwen3-8B 评估）之间的相关性。所有结果均具有统计显著性，$p < 0.001$。

具体而言，Qwen3-8B（Yang 等人，2025年a）独立评估模型预测结果，同时由一个由 2–3 名具备相关领域专业知识的硕士生和博士生组成的评审小组进行人工评估。我们通过多数投票汇总人类评分，以确立评估黄金标准。随后，我们计算了自动评分与这一人类共识之间的相关性。如表 7 所示，自动评判器与黄金标准表现出高度一致（平均皮尔逊 $r = 0.825$），证实 Qwen3-8B 可作为人类评估的有效且一致的替代物。

E 补充实验与结果分析

E.1 学科层面的性能分析

图 10总结了我们一级分类体系下各学科的表现。我们观察到不同学科之间的差异，表明 SIN-Bench 捕捉到了基于证据的科学理解方面有意义的变化。综合各模型来看，经济学与金融学和 医学与健康科学 整体表现相对较强（平均 $\approx$ 0.66 和0.65），而 数学

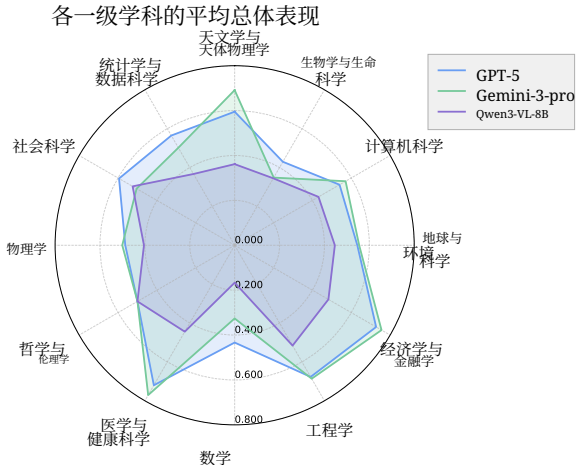


图10：SIN-Bench 各领域的表现。雷达图展示了 GPT-5、Gemini-3-pro 和 Qwen3-VL-8B 在 12 个主要学科中各项任务的总体得分平均值。得分越高越好，图中凸显了基于证据的科学理解在不同领域存在显著差异。

始终是最具挑战性的领域（平均 ≈ 0.31）。这种对比表明，严谨的符号运算和定量依据仍是当前多模态大语言模型面临的主要瓶颈。

模型排名也因学科而异。Gemini-3-pro 在视觉信息密集或布局复杂的学科中表现出明显优势，例如天文学与天体物理学和医学与健康科学（例如，$+0.01$ 和 $+0.05$ 优于 GPT-5），这与其在证据主要通过图、表格和图表类图形传达时具有更强的跨模态锚定能力相符。相比之下，GPT-5 在更偏形式化或定量的学科中表现更好，其中最突出的是数学（$+0.10$优于 Gemini-3-pro）和统计学与数据科学。Qwen3-VL-8B 在大多数学科中落后于前沿模型，但在社会科学 中仍具备相对竞争力，并在哲学与伦理学中与其他模型表现趋同；在这些领域，判断可能更多依赖一般性话语理解，而较少依赖细粒度的多模态证据。

E.2 SIN-Verify 中近失证据的分析

为探究超越简单不匹配的验证难度，我们在交叉验证过程中构造了两组负例集（第A.2节，步骤（2））。简单负例将问题与明显无关的证据配对，因此很容易予以驳回。困难负例则是近似匹配的案例：问题质量高，答案也正确，但证据链未能通过交叉验证的一致性检验，也就是说，证据存在歧

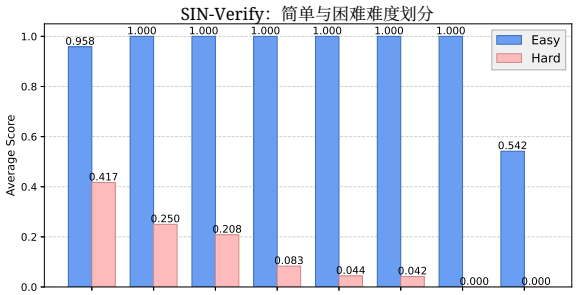


图11：在 SIN-Verify 的两个负例集上的准确率。简单负例是明显无关的证据；困难负例则是经交叉验证确认的近失案例，答案正确，但证据支持不足或含糊不清。

义或不足以确凿地支持答案。每种设置下我们抽取 24 个实例。

如图 11所示，多模态大语言模型在简单负例上的表现几近饱和，表明其具备很强的粗粒度不匹配检测能力。相比之下，在困难负例上的准确率急剧下降，说明近失证据使验证比相关性检测难得多。模型必须通过细粒度的跨模态定位（例如数字、条件以及图文对齐）来评估证据是否充分，同时避免答案驱动的合理化。Gemini-2.5-pro 在困难负例上表现最佳（0.417），这与其在判定“证据不足”方面校准能力更强的情况一致。Gemini-3-pro（0.250）和 GPT-5（0.208）仍远未达到上限，这表明，面对近失证据时，整体推理能力并不会直接转化为更严格的证据支持判断。其余模型的表现则降至接近零，这与受表面相关性线索主导的验证行为相符。

E.3 文档长度对性能稳定性的影响。

科学论文从其构成上看，篇幅较长且具有多模态特征。为刻画交错输入下性能随上下文长度的变化情况，我们使用两种词元计数方案分析提示长度与总体得分之间的联合分布：纯文本词元（图 12）和文本 + 图像词元（图 13）。本小节报告 SIN-Find、SIN-QA 和 SIN-Summary 的结果。在第 5.3节中，我们将进一步聚焦于对 SIN-QA 和 SIN-Summary 的解读，因为这两个任务属于高级任务，需要对文档进行整体理解，并进行基于证据的生成。

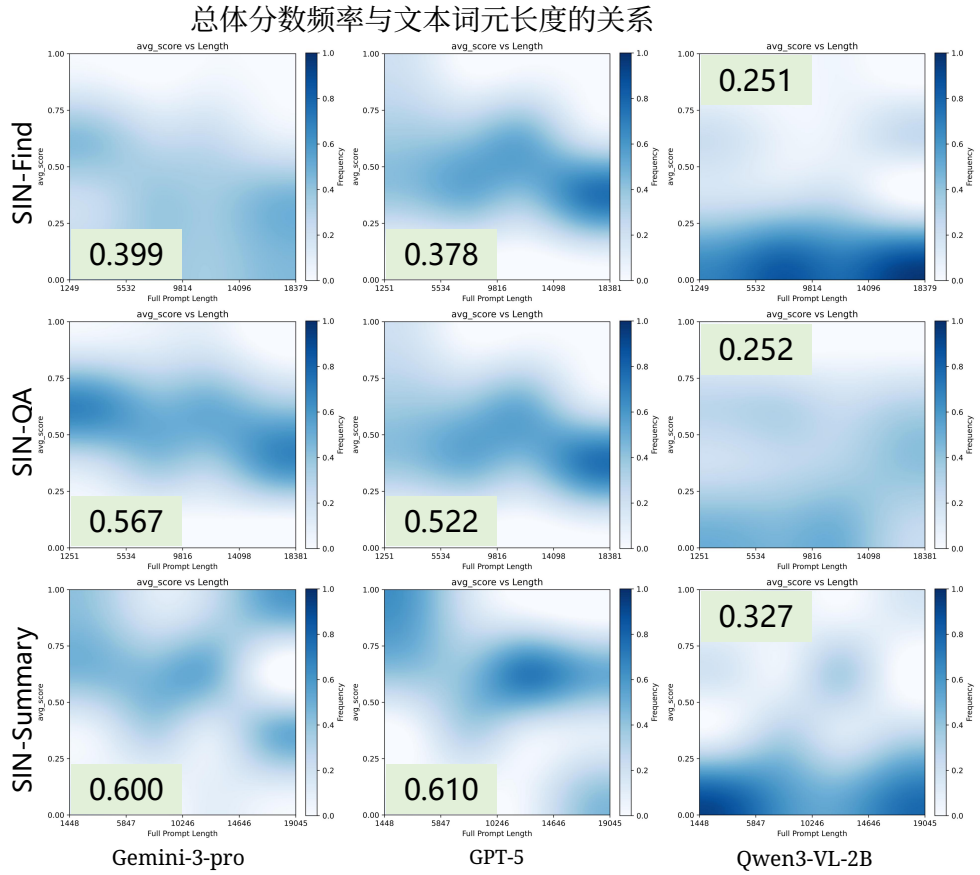


图 12：SIN-Find、SIN-QA 和 SIN-Summary 的总体得分频次与文本词元长度的关系。按文本提示长度和总体得分分箱的样本二维频次归一化结果。

设置。对于每个经过评估的样本，我们记录应用程序接口调用中使用的真实输入词元数，以及该样本的总体得分。在图 12中，x 轴仅统计文本词元。我们将此视图作为主要的公平性轴，因为不同模型的英语分词大体相似，而图像分词可能高度依赖具体模型和流水线。因此，我们认为，仅使用文本词元长度能够为跨模型比较提供更具代表性且更公平的依据。在图 13中，x 轴统计文本+图像词元，其中图像词元取决于视觉编码策略（例如，原生分辨率输入与调整大小后的输入）。对于 Qwen3-VL-2B，词元统计遵循官方配置。每个面板展示（提示长度、总体得分）上的归一化频率；插图报告总体得分的均值。

纯文本长度（图12）：长文本上下文下的稳健性。在全部三个任务中，Gemini-3- pro 和 GPT-5 在完整文本长度范围（约1k–19k 个词元）内的中高分区域均占据了相当大的概率质量。随着提示词变长，GPT-5 的得分下降幅度比 Gemini-3-pro 更为明显。然而，在最长上下文中，两个模型都没有表现出概率质量系统性地坍缩到低分区间的现象。这一模式表明，对于强大的模型而言，仅文本长度增加不太可能是主要瓶颈；剩余错误更可能源于科学叙述中的选择性证据识别和细粒度定位。

相比之下，Qwen3-VL-2B 始终处于低分区间。对于 SIN-Find 和 SIN-QA，得分分布集中在低分段，即使文本长度较短，得分也很少向高分段迁移。SIN-Summary 也呈现出类似的偏斜，整体表现始终以低分为主。总体而言，这些趋势表明，Qwen3-VL-2B 的瓶颈不仅在于长上下文处理，更在于交错式科学输入下稳健证据检索、跨模态对齐和高层次综合能力方面存在更广泛的差距。

文本+图像长度（图13）：视觉标记化导致的长度扩展和聚类。纳入图像标记后，提示长度

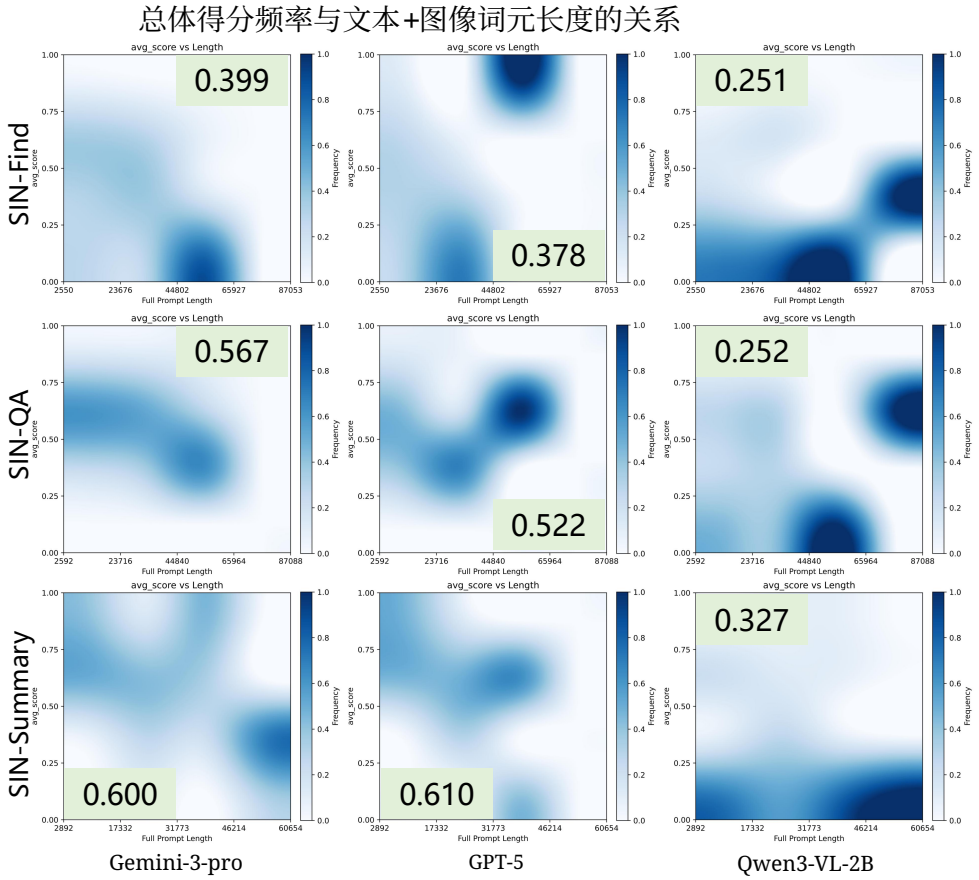


图 13: 总体得分频率与总（文本和图像）词元长度的关系。与图 12 相同，但提示长度为 th 计算得出, 依据 API 请求中的文本和图像词元计算。

范围大幅扩展（最多达到 ∼87k 个词元）。这种扩展主要反映了图数量以及特定于模型的视觉编码。与图12相比，分布沿 x 轴也变得更加聚集。这与总长度离散成由典型图数量/分辨率及底层视觉标记器所决定的不同区间相一致。因此，应谨慎解读总长度轴上的趋势，因为它混合了文档的视觉内容与预处理和编码选择的影响。

Gemini-3-pro 和 GPT-5 的平均总体得分相近，但随着上下文增长，两者的分布在定性上有所不同。Gemini-3-pro 的得分分布相对平稳，主要落在中高分区间，并且在不同长度下趋势一致。相比之下，GPT-5 呈现出更明显的两极分化：随着总长度增加，概率质量分成两部分，一种模式仍处于高分区域，另一种则向低分区域移动。一种可能的解释是，总长度的增加主要来自额外的图，从而加大了视觉落地难度的方差。当新增的图具有冗余性或能直接提供支持时，GPT-5 可以利用额外的多模态证据并保持高分。然而，当较长的输入引入视觉相似的干扰项、稠密的多图证据链，或分辨率敏感线索（例如小字、数字、坐标轴标签）时，出错的可能性会增加，从而形成低分模式。换言之，较长的多模态上下文并不会一概降低性能；相反，它们会混合“有帮助”和“对抗性”的视觉补充，从而增加实例级方差。

对于 Qwen3-VL-2B，各项任务的得分分布仍主要集中在低分区域。不过，在 SIN-Find 和 SIN-QA 中，随着总长度增加，中等分数区间的样本占比也不可忽视。一个可能的原因是，这些较长的提示词增加的主要是图像数量，而非文本长度；只要包含正确的图表，有时就能凭借显著的视觉线索或近似的表层匹配完成任务。相比之下，即使总长度很大，SIN-Summary 的得分仍几乎全部集中在低分区域，这表明增加视觉词元

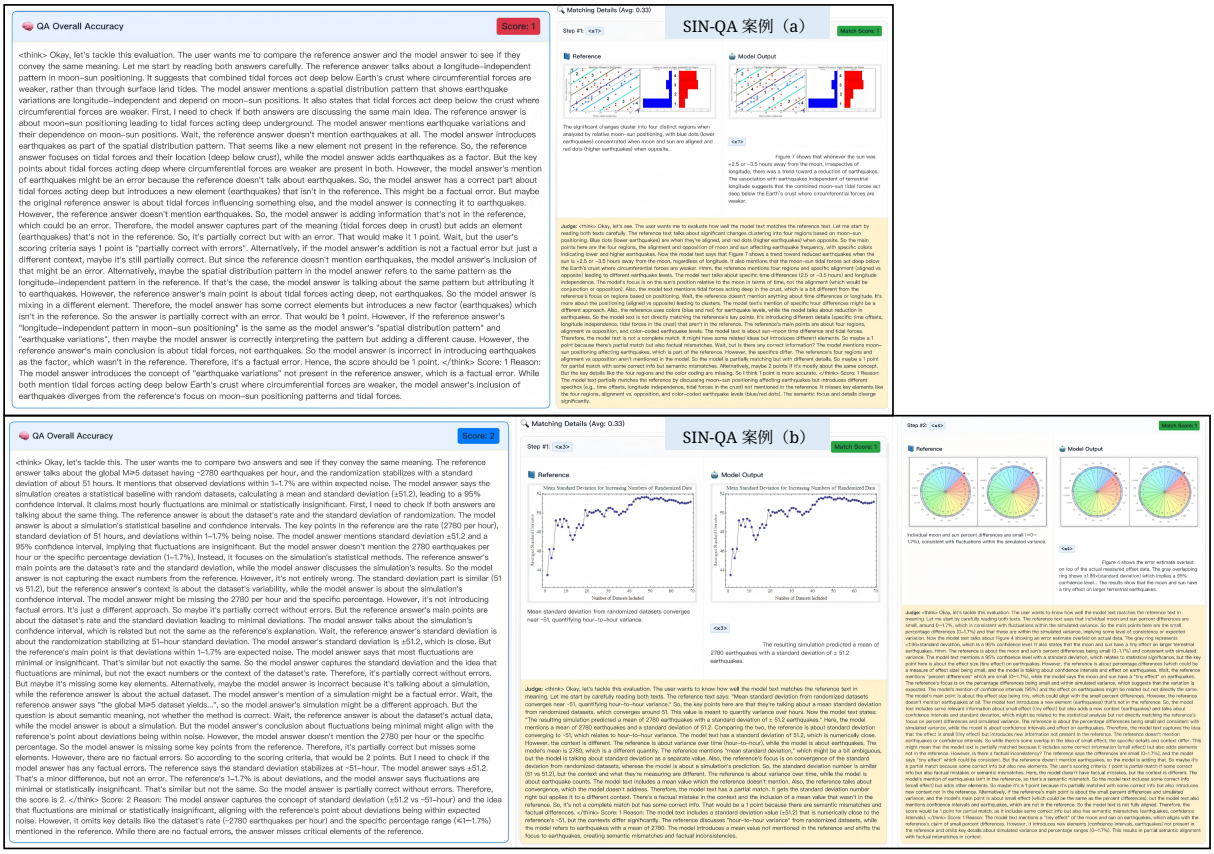


图14：SIN-QA 上的错误案例。

本身并不能缓解核心难点：在整个文档层面组织并综合证据，从而形成连贯的摘要。

导致结果看似定量，却违反了关键的方法学约束，进而得出与方法不一致的结论。

F 案例分析

如图 14 所示，我们进一步分析 SIN-QA，因为它既要求给出最终答案，也要求提供明确的证据链，因此适合用于诊断总体得分之外的依据失败。除了第 5.3 节总结的两种失败模式（信息不足和虚假推理）之外，我们的案例分析还揭示了两种具体表现形式。

在案例 (a) 中，模型引入了参考内容中不存在的无依据概念（例如，将潮汐力解释扩展成地震叙事）。尽管部分陈述表面上仍有关联，但新增内容改变了解释框架并导致语义漂移，体现了**虚假推理**：流畅且自信的输出受到缺乏依据的扩展内容污染，降低了精确率。相比之下，案例 (b) 体现了**信息不足**的一种程序性表现：模型未遵循参考内容预期的统计解释，而是混淆或过度简化了标准差不变性分析的作用，