大语言模型、人工审稿人与作者在 Ev风湿病学观察性研究 STROBE 清单的填写

这篇手稿为预印本，尚未经过同行评审。

Emre Bilgin[1], Ebru Öztürk[2], Meera Shah[3], Lisa Traboco[4,5], Rebecca Everitt[6], Ai Lyn Tan[7], Marwan Bukhari[8,9], Vincenzo Venerito*[10], LaDka Gupta*[11,12,13]

[1] 土耳其萨卡里亚，萨卡里亚大学医学院内科风湿病科[2] 土耳其安卡拉，哈杰泰佩大学医学院生物统计学系[3] 印度孟买，托皮瓦拉国立医学院及 B.Y.L. Nair医院风湿病科[4] 菲律宾圣卢克医疗中心（环球城）[5] 菲律宾马尼拉，菲律宾大学（医学信息学部）[6] 英国伍尔弗汉普顿，新十字医院，皇家伍尔弗汉普顿信托[7] 英国利兹，NIHR利兹生物医学研究中心、利兹教学医院NHS信托；利兹大学利兹风湿病与肌肉骨骼医学研究所[8] 英国兰卡斯特，兰卡斯特大学医学院，兰卡斯特大学[9] 英国兰卡斯特，皇家兰卡斯特医院风湿病科[10] 意大利巴里，巴里阿尔多·莫罗大学精准与再生医学及爱奥尼亚地区系（DiMePReJ）风湿病科，Umberto I广场，70121[11] 英国伍尔弗汉普顿，皇家伍尔弗汉普顿医院NHS信托风湿病科。[12] 英国伯明翰，伯明翰大学医学与健康学院感染、炎症与免疫学系。[13] 英国伦敦，弗朗西斯·克里克研究所。*共同资深作者

开放研究者与贡献者身份标识：

Emre Bilgin：0000-0002-2260-4660 Ebru Öztürk：0000-0001-9206-6856 Meera Shah：0009-0005-3308-7231 Lisa Traboco：0000-0002-1952-7879 Rebecca Everitt：0000-0003-1919-5293 Ai Lyn Tan：0000-0002-9158-7243 Marwan Bukhari：0000-0003-4311-5222 Vincenzo Venerito：0000-0002-2573-5930 LaDka Gupta：0000-0003-2753-2990

通信作者：
Vincenzo Venerito，巴里阿尔多·莫罗大学精确与再生医学及爱奥尼亚地区系（DiMePReJ）风湿病科，Piazza Umberto I，70121，意大利巴里，vincenzo.venerito@gmail.com

要点：

- 在我们的样本中，大语言模型在基本的 STROBE 格式方面与人工审稿人高度一致，表明其可能适用于初步筛查。- 然而，大语言模型对复杂的方法学项目的评估可能与资深人类专家的评估不同。- 尽管模型有助于进行结构检查，但评估方法学仍离不开人类专家判断。

摘要

引言： 评估对《加强流行病学观察性研究报告规范》（STROBE）声明的遵循情况可能耗时且带有主观性。本研究比较了大语言模型（LLMs）、人类评审小组和原稿作者对风湿病学观察性研究的STROBE评估结果。

方法： 本研究以《可靠性与一致性研究报告指南》（GRRAS）和DEAL路径B框架为指导，对17篇风湿病学论文进行了独立评估。评估采用包含22项的STROBE清单，由作者、五人组成的人类评审小组（成员从初级到高级专业人员不等）以及两个大语言模型（ChatGPT-5.2、Gemini-3Pro）分别完成。条目分为“方法学严谨性”和“呈现与背景”两个领域。评审者间可靠性采用Gwet一致性系数（AC1）计算。

结果： 所有审查者之间的总体一致率为 85.0%（AC1=0.826）。领域分层显示，“呈现与背景”几乎达到完全一致（AC1=0.841），而“方法学严谨性”达到实质性一致（AC1=0.803）。尽管大语言模型在标准格式元素上与所有人工审稿人达成完全一致（AC1=1.000），但在复杂条目上，其与人工审稿人及作者的一致性有所下降。例如，对于失访条目，Gemini 3 Pro 与高级审稿人之间的一致性为 AC1=-0.252，而与作者之间仅达到一般一致。此外，在特定方法学条目上，ChatGPT-5.2 通常比 Gemini- 3Pro 与人工审稿人表现出更高的一致性。

结论： 尽管大语言模型在基础 STROBE 筛查方面展现出潜力，但其在复杂的方法学项目上与人类专家的一致性较低，这可能反映出其依赖表层信息。目前，与其取代人类专家判断来评估观察性研究，这些模型似乎更适合用于规范化处理直观易懂的检查。

关键词： 大语言模型、STROBE清单、评定者间信度、观察性研究、风湿病学

引言

自早期模型（如生成式预训练变换模型 3（GPT-3））问世以来，大语言模型（LLMs）已迅速成为学术研究的一部分 [1]。它们越来越多地用于文献筛选、数据提取和代码生成等任务，有助于减少人工工作量和人为错误 [2]。然而，人工智能在关键科学流程中的应用仍存在争议，主要原因是可靠性、偏倚和透明度方面的问题尚未解决。 [3]。

透明报告对于评估观察性研究至关重要。加强流行病学观察性研究报告（STROBE）声明提供了一份包含 22 个条目的核对清单，以标准化这一流程并提高结果的可解释性 [4]。然而，在实践中，评估 STROBE 依从性既耗时，又容易受到评审者间变异的影响，因为缺乏严格客观的标准 [5]。此外，正如 Sharp 等人指出的，许多作者将核对清单视为耗时的行政要求，而非实用工具，这可能会影响自我报告依从性的准确性 [5]。

本研究旨在系统评估当前大语言模型在评估风湿病学观察性研究对 STROBE清单的依从性方面的表现。通过将其表现与临床专业水平各异的人工审稿人小组以及原稿作者直接比较，我们试图探究人工智能（AI）作为辅助工具的潜力，以标准化报告评估，并帮助读者和研究人员批判性地评估观察性文献。

方法

研究设计、评审者间一致性分析及报告框架遵循《可靠性与一致性研究报告指南》（GRRAS）的原则，并与《大语言模型的开发、评估与评价》（DEAL）路径 B 核对清单保持一致，用于预训练模型的应用评估 [6, 7]。

搜索策略与筛选标准：

我们利用国际研究 Observatory（OOIR） [8]，整理了一份按影响因子排名的顶级风

通过既往文献计量分析及相应的《作者指南》[9, 10]进行核实。在入选期刊中，我们通过结构化检索确定了所有于2020年2月至2025年2月期间发表且提及“STROBE”的文章。我们纳入了17篇包含完整且可获取的 STROBE 清单的文章进行评估（完整参考文献见**补充文本-1**）。筛选流程详见研究筛选流程图（**补充图- 1**）。

评估方案**：**

由三名初级、一名中级和一名高级专业人员组成的人工审稿人小组，使用 STROBE 清单对这17篇文章进行了独立评估。此外，两个人工智能大语言模型——ChatGPT-5.2 和 Gemini 3 Pro——也进行了评估。为便于比较，我们还收集了作者最初自行报告的 STROBE 清单。

为确保结果可复现并遵循 DEAL-B框架，评估通过两种广泛使用的专有闭源模型的官方网页界面，使用其高级订阅版本进行：ChatGPT-5.2（自动模式；OpenAI）和 Gemini 3 Pro（Google有限责任公司）。所有测试均于2026年1月10日进行。选择网页界面而非 API，是为了模拟真实的临床环境；温度和 top-p 等生成参数均保持供应商默认设置。每份手稿都在新的聊天会话中进行评估，以确保评估相互独立。我们采用了三步提示方法（完整提示词见**补充文本-2**）。具体包括：(1) 指示模型仅评估报告规范的遵循情况，而非方法学质量，使用严格的 Y/N/NA 分类，并告知模型不要推断缺失的细节；(2) 通过两篇附有人工作为参考评分的示例文章进行少样本校准，以设定报告充分性的判定阈值；(3) 进行最终评估，使用不带解释或推理的简单两列表格，以防止模型生成虚构的解释。所有模型输出均按生成时的原样记录，未经人工修改。

评分与统计分析：

所有审稿人（人工和机器）均将每个 STROBE 条目归类为“是”（报告充分）、“否”（报告不充分或未报告）或“不适用”。在开始数据收集之前，我们的研究团队将 STROBE 条目分为两类：“方法学严谨性”

（条目 4–12）和“呈现与背景”（条目 1–3、13–22）。这与原始 STROBE 清单一致，其中条目 4–12 构成“方法”部分。我们计划采用这种分类，以考察模型在评估核心方法学步骤和一般报告实践时的表现是否不同。我们评估的是评定者间信度，而非诊断准确性（如敏感度），因为评估报告指南本质上依赖主观解读，且不存在绝对的“真实标准”。我们采用总体一致率和 Gwet 首次评定者间一致性系数（AC1）及其 95% 置信区间（CI），并使用 “irrCAC” 软件包[11]，量化人工审稿人、大语言模型与原稿作者之间的评定者间信度。选择 Gwet 的 AC1 来估计可靠性，是因为它能够校正偶然一致，同时避免高度偏斜的核对清单数据集中经常出现的 Cohen κ 患病率悖论。可靠性估计采用标准阈值进行解释：<0.00（较差）、0.00–0.20（轻微）、0.21–0.40（一般）、0.41–0.60（中等）、0.61–0.80（较高）和 0.81–1.00（近乎完美） [12]。所有评估不同评定者组合间总体、领域特定及条目特定一致性的统计分析，均使用 R 软件 4.4.3 版（R统计计算基金会，奥地利维也纳）完成。

结果

总体及特定领域的一致性

在筛查的189篇手稿中，只有17篇文章提供了完整且可获取、适用于分析的STROBE清单（**补充图1**）。所有条目和审稿人（人工评审、大语言模型、作者）之间的总体一致率为85.0%（Gwet’s AC1 = 0.826 [95% CI 0.801– 0.851]），表明一致性几乎完美。按领域分层后发现，“呈现与背景”的一致性几乎完美（AC1 = 0.841 [95% CI 0.810–0.872]），“方法学严谨性”的一致性较高（AC1 = 0.803 [95% CI 0.761–0.845]），所有 p 值均为 <0.001。两两分析显示，所有评估组之间的评定者间信度高度一致，绝大多数组合均表现出几乎完美的一致性（AC1 > 0.80；图1和图2）。在文章层面，总体 AC1 介于0.624至0.931之间；与“呈现与背景”（0.685–0.946）相比，“方法学严谨性”领域的变异更大（0.532–0.921）。

条目级分析

对各项STROBE条目及子条目一致性评分的考察发现，尽管审稿人在许多部分达成了近乎完美的一致，但核对清单各项的吻合程度自然有所不同。某些具体问题上出现了明显差异，有时评分甚至显著偏低。补充图 2 展示了所有评估组针对各具体条目的详细两两比较结果。

- **标准报告条目的一致性较高**

在“呈现与背景”领域中，两种语言模型均持续识别出手稿的标准章节。对于基本报告要素——如信息丰富的摘要（条目 1b）、背景与研究依据（条目 2）、具体目标（条目 3）、关键结果摘要（条目 18）、研究局限性（条目 19）和总体解释（条目 20）——ChatGPT- 5.2 和 Gemini 3 Pro 与所有人工审稿人及原作者均达成完全一致（AC1 = 1.000）。在识别资金来源（条目 22）和参与者人数（条目 14a）方面，一致性也同样很高。

- **复杂方法学条目的差异**

条目级分析显示，差异更为明显，尤其体现在方法学报告方面，具体来说是第12项。对于更复杂的条目，大语言模型与高级审稿人之间的一致性有所下降。例如，在失访（条目 12d）方面，大语言模型之间的一致性为中等（AC1 = 0.447），两个模型与作者和初级审稿人之间也均表现出一般一致性。然而，它们与高级审稿人之间的一致性显著下降；Gemini 3 Pro 与高级审稿人之间的一致性落入负值范围（AC1 = - 0.252）。对于敏感性分析（第12e项），大语言模型彼此之间具有较高一致性（AC1 =0.652），与高级审稿人之间的一致性为中等，但与作者之间仅表现出轻微至一般的一致性（AC1: 0.071–0.272）。最后，在报告连续变量分类（条目 16b）方面，大语言模型彼此之间的一致性为中等（AC1 = 0.564），人工组之间总体上也较为一致；然而，大语言模型与人工审稿人之间的一致性属于一般和轻微级别。

- **不同语言模型之间的表现差异**

两种语言模型之间的一致性水平也存在差异。在评估缺失数据的处理方式（第12c项）时，尽管大语言模型之间具有较高一致性（AC1 = 0.668），初级审稿人与两个模型的一致性均较低，而且与人工审稿人相比，Gemini 3 Pro 通常表现出低于 ChatGPT-5.2 的一致性。作者在该条目上与初级审稿人保持了较高一致性。在报告缺失数据参与者人数（条目 14b）方面，模型表现各异。ChatGPT-5.2 与作者（AC1 = 0.730）及中级审稿人（AC1= 0.722）均达到较高一致性，而 Gemini 3 Pro 则分别表现出一般（AC1 = 0.301）和中等（AC1 = 0.450）一致性。

讨论

在本研究中，我们考察了大语言模型、人工审稿人小组与稿件作者在评估风湿病学观察性研究对STROBE清单的遵循情况时，彼此之间的一致程度。尽管总体一致率为85.0%，但这主要由“呈现与背景”领域推动。该部分的条目通常要求识别标准的结构组成部分，这与当前语言模型基于文本匹配的基本特性十分契合 [13]。相比之下，“方法学严谨性”领域的一致性较低且差异更大，这表明对于所有类型的审稿人而言，评估复杂的方法学报告仍然具有挑战性。

条目级结果表明，审稿人的背景可能会影响其评分模式。关于敏感性分析的报告（第12e项），作者通常与初级审稿人意见一致，但与大语言模型仅达到轻微至一般程度的一致。模型的评分与高级审稿人的评分更为接近。这一发现与现有研究一致，即作者往往比独立审稿人更积极地看待自己报告的方法学内容 [14]。在我们的样本中，初级审稿人似乎更倾向于接受作者的陈述，而资深专家和语言模型则倾向于采用更严格的标准。因此，这些发现表明，纳入大语言模型可能有助于通过补充性核查来验证作者报告的清单。

我们还观察到，模型评估结果有时会偏离人工审稿人的判断。对于处理失访（条目 12d）和连续变量分类（条目 16b）等任务，模型与高级审稿人之间的一致性降至轻微或较差水平。这可能是因为模型高度依赖字面表述。如果论文没有明确说明如何处理失访，模型往往会将其视为缺失数据。相比之下，经验丰富的人工审稿人通常可以从数据表或流程图中推断出这一信息。这是一个已知的局限：模型通常能够较好地处理表层文本，但难以识别隐含的方法学细节 [15]。未来研究应对这些不一致案例开展定性分析，以更好地理解产生这些分歧的具体原因。

除了人工审稿人与模型输出之间的差异外，两个语言模型之间也存在性能差距。例如，在评估缺失数据计数（条目 14b）时，ChatGPT-5.2 与人工审稿人的结果高度一致，而 Gemini 3 Pro 的一致性评分较低。此前针对医学基准测试的研究同样表明，输出可能会随模型的训练与更新周期而变化 [16]。这凸显了研究人员谨慎选择模型的重要性，因为所选工具可能会显著影响评估结果。

本研究有几项重要优势。首先，我们组建了一个多元化的人工审稿人小组，其中既有初级研究人员，也有资深专家，同时还纳入了原稿作者。这样的设置使我们能够直接观察临床经验如何影响核对清单评分。其次，我们采用标准化的独立提示词流程测试了两种语言模型，全程不进行人工干预，从而确保机器评估具有可重复性。第三，我们使用 Gwet 的 AC1 统计量衡量一致性，这一指标可避免核对清单数据中使用 Cohen κ 时常见的患病率悖论，从而更准确地估计评定者间信度。然而，本研究也存在局限性。首要局限是样本量较小，仅纳入 17 篇文章，这自然会降低统计效能并扩大置信区间，尤其是条目层面分析的置信区间。不过，这一数量并非方法学上的选择，而是当前文献现状的直接反映。尽管我们筛查了大量风湿病学观察性研究，但仅发现 17 篇明确报告使用 STROBE 清单的研究。虽然这一小样本使我们的估计稳定性较低，但也揭示了一个更广泛且关键的问题：

本领域普遍缺乏对这一指南的报告。此外，由于我们的样本由来自高影响力期刊、且明确要求使用 STROBE 清单的文章构成，我们评估的报告质量可能高于平均水平。在更广泛的出版物范围内，大语言模型与人工审稿人之间的一致性可能有所不同。在分析方面，将 STROBE 清单分为两个领域，有助于我们区分方法与格式。不过，我们也认识到这种划分并不完美；例如，条目 16b 属于呈现领域，但其一致性评分较低，表明它更像一项复杂的方法学任务。我们也只评估了两种商业大语言模型，未纳入 Llama 和 Mistral 等开源模型。我们之所以这样选择，是为了反映研究人员的日常实际情况：他们通常依赖标准、即用型网页界面，而非需要特定硬件的本地安装。未来研究应探讨开源模型或专用医学模型在这些任务中的表现。我们还选择每个模型对每篇文章仅评估一次，因此未评估评定者内一致性。语言模型对同一提示词可能生成不同的回答，多次运行可以显示其答案的波动程度。不过，我们有意采用单次评估方式，以模拟人工审稿人的实际工作流程；他们通常只阅读并评分一次手稿。允许多次运行后再选择最准确的输出，也会引入选择偏倚。对于未来的应用，整合检索增强生成（RAG）系统是解决这种固有变异性的实用方案 [17]。通过将模型锚定于专门检索到的文本片段，RAG 架构能够显著提高自动化核对清单评估的稳定性和可靠性 [17]。此外，还存在固有的数据泄漏风险。由于接受评估的文章已经公开，它们很可能已被纳入模型的训练数据。模型可能依赖先前接触过这些文章，而不是仅从我们提供的文本中提取信息，这可能会抬高一致性率。即使严格要求模型仅使用所提供的文本，也无法将大语言模型与其预训练记忆完全分离。除了预训练记忆之外，模型本身也在不断演进；由于我们是在特定时间点测试 ChatGPT-5.2 和 Gemini 3 Pro，未来的更新自然可能产生不同的结果。

总之，语言模型有望成为评估一般报告标准的实用工具，因此适用于基础的 STROBE 筛查。然而，它们与高级

人类专家在复杂的方法学项目上的一致性较低，这可能反映出它们依赖表层信息，而非深入理解方法学。虽然这些模型有助于规范简单的核对清单任务，但在准确评估观察性研究中的研究方法时，人类专家判断仍不可或缺。

补充材料见本手稿末尾。

伦理审批

本研究无需进行伦理审批，因为本研究仅涉及对既往发表且公开可获取的观察性研究进行方法学分析，未涉及人类或动物受试者。

数据可用性声明

本研究期间生成和分析的数据集，包括所有人类评分者的去标识化评分表以及语言模型输出的完整逐字稿，可根据合理请求向通信作者获取。

经费

作者未从公共、商业或非营利部门的任何资助机构获得用于本研究、作者工作和/或本文发表的专项资助。

人工智能（AI）披露

在准备本手稿期间，作者仅使用 Gemini 3 Pro（Google）进行语言编辑、语法润色和可读性改进。使用该工具后，所有作者均对手稿进行了全面审阅和编辑，并批准了最终版本。作者对本文的完整性、准确性和原创内容承担全部最终责任。

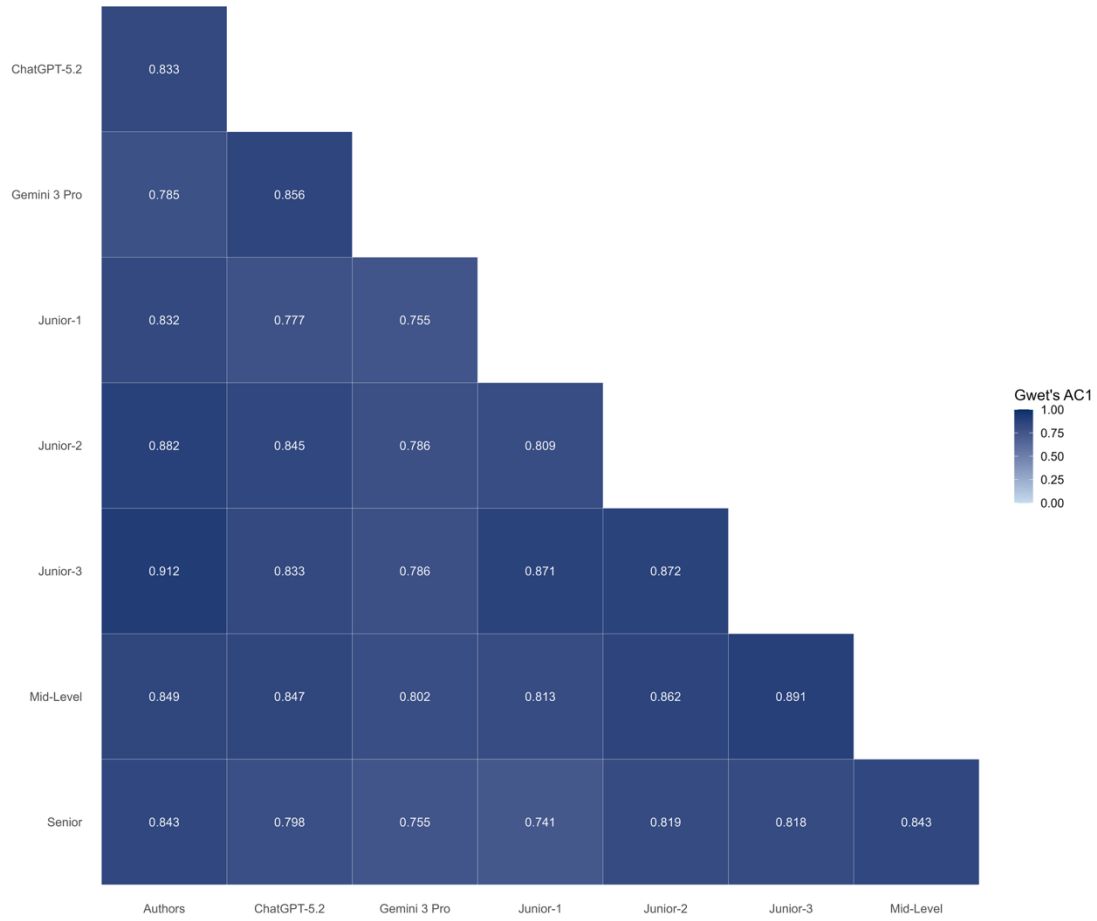


图 1. **审稿人之间的总体两两一致性**。 热图展示了 Gwet 首次一致性系数（AC1）的估计值，反映所有 STROBE 清单条目上的总体两两比较结果。审稿人包括原稿作者、两个大语言模型（ChatGPT-5.2 和 Gemini 3 Pro），以及由独立人工审稿人组成的小组（三名初级、一名中级和一名高级审稿人）。

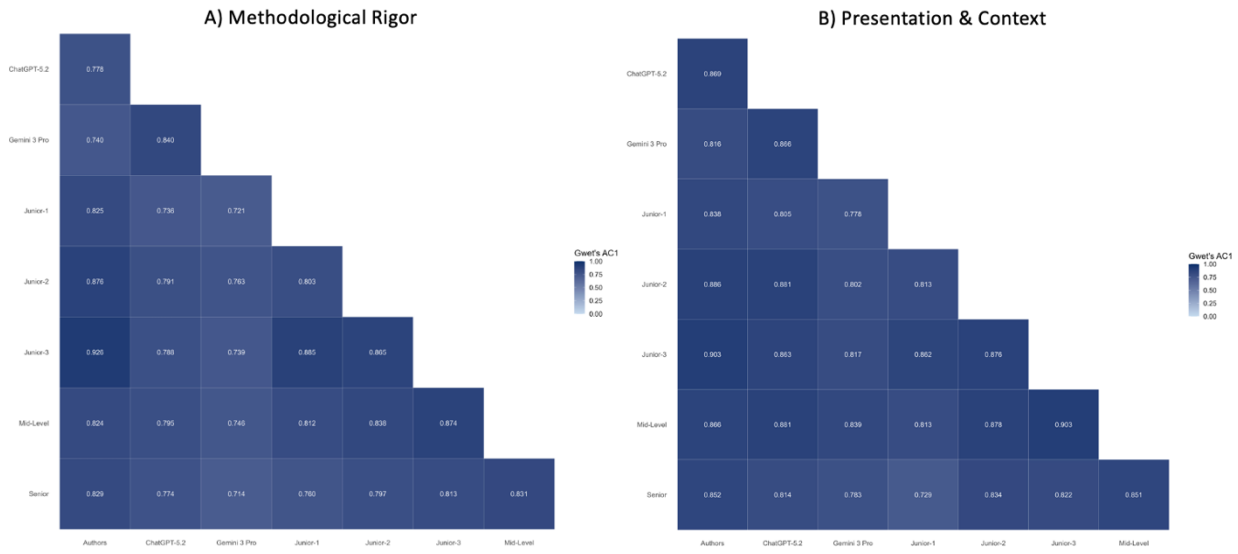


图 2. 按 STROBE 领域分层的两两一致性。热图展示了针对两个不同评估领域的两两比较所估算的 Gwet 首阶一致性系数（AC1）：(A) 方法学严谨性（条目 4–12）和 (B) 呈现与背景（条目 1–3、13–22）。审稿人包括原稿作者、两个大语言模型（ChatGPT-5.2 和 Gemini 3 Pro），以及由独立人工审稿人组成的小组（三名初级审稿人、一名中级审稿人和一名高级审稿人）。

参考文献：

1. Venerito, V. and L. Gupta，大语言模型：风湿病学家最新的同事？*Nat Rev Rheumatol, 2024. 20(2): p. 75-76. 2. Venerito, V., et al.,* 我是人工智能风湿病学家：面向风湿病学家的大语言模型实用入门。*Rheumatology (Oxford), 2023. 62(10): p. 3256-3260. 3. Venerito, V., et al.,* 人工智能在风湿病学中的伦理挑战与监管路径。*Rheumatol Adv Pract, 2025. 9(2): p. rkaf035. 4. von Elm, E., et al.,* 加强流行病学观察性研究报告规范（*STROBE*）声明：观察性研究报告指南。*J Clin Epidemiol, 2008. 61(4): p. 344-9. 5. Sharp, M.K., K. GlonD, and D. Hren*，关于 *STROBE* 声明的在线调查凸显了对其内容、目的和价值的不同看法。*J Clin Epidemiol, 2020. 123: p. 100-106. 6. Kottner, J., et al.,* 《可靠性与一致性研究报告指南》（*GRRAS*）提案。*J Clin Epidemiol, 2011. 64(1): p. 96-106. 7. Tripathi, S., et al.,* 大语言模型的开发、评估与评价（*DEAL*）核对清单：一份技术报告。*NEJM AI, 2025. 2(6): p. AIp2401106. 8. Observatory, et al.;* 网址：*https://ooir.org/*。 *9. Sharp, M.K., et al.,* 一项横断面文献计量学研究显示，期刊对 *STROBE* 及其扩展版本的认可率不尽理想。*J Clin Epidemiol, 2019. 107: p. 42-50. 10. Barajas-Ochoa, A., et al.,* 风湿病学中报告指南的使用：对发表于*5*本主要风湿病学期刊的*850*篇以上稿件开展的横断面研究。*J Rheumatol, 2023. 50(7): p. 939-943. 11. Gwet, K.L. irrCAC*：计算机会校正一致性系数（*CAC*）。*2019*；网址：*https://CRAN.R-project.org/package=irrCAC*。 *12. Gwet, K.L.*，《评分者间信度手册：衡量评分者间一致程度的权威指南》。*2014*，马里兰州，美国：马里兰州盖瑟斯堡：*Advanced AnalyDcs, LLC*。 *13. Garcia-Carmona, A.M., et al.,* 利用大语言模型从医疗报告中准确检索患者信息：系统评估研究。*Jmir ai, 2025. 4: p. e68776. 14. Rastogi, C., et al.,* 作者对自己论文的看法与合著者的看法及同行评审决定有何异同？*PLoS One, 2024. 19(4): p. e0300710. 15. Paci, W., A. Panunzi, and S. Pezzelle.* 他们想要假装不理解：当前大语言模型解读政治话语隐含内容的局限。载于《计算语言学协会论文集：*ACL 2025*》。*2025*。*16. NDnopoulos, V., et al.,* 利用大语言模型从非结构化和半结构化电子健康记录中提取数据：多模型性能评估。*BMJ Health & Care InformaDcs, 2025. 32(1): p. e101139. 17. Li, Z., et al., RAPID*：利用大语言模型可靠、高效地自动生成投稿报告清单。*Journal of the American Medical InformaDcs AssociaDon, 2025. 32(8): p. 1340-1349.*

补充材料

补充文本-1：纳入文章列表。 对针对 STROBE清单一致性进行评估的17项观察性研究的完整参考文献

1. Larid, G. 等，风湿病学和内科学科室住院成人中低磷酸酯酶症的检出：一项为期*10*年的多中心研究。 RMD Open，2024年。**10**（2）。2. Bastard, L. 等，银屑病关节炎患者使用不同类别靶向治疗时严重感染的风险：基于法国健康保险数据库（*SNDS*）的全国队列研究。 RMD Open，2024年。**10**（1）。3. Sloan, M. 等，系统性自身免疫性风湿病中神经精神症状的患病率和识别：一项国际混合方法研究。 Rheumatology (Oxford)，2024年。**63**（5）：第1259-1272页。4. Crossfield, S.S.R. 等，二十年来类风湿关节炎药物治疗管理的变化。 Rheumatology (Oxford)，2021年。**60**（9）：第4141-4151页。5. Spierings, J. 等，系统性硬化症患者如何看待当前提供的医疗保健服务，以及我们应如何衡量其质量？ Rheumatology (Oxford)，2020。**59**（6）：第1226-1232页。6. Garriga, C. 等，急性膝关节损伤后*2*年结局的临床和分子关联：肯尼迪膝关节损伤队列（*KICK*）中的一项纵向研究。Lancet Rheumatol，2021年。**3**（9）：第e648-e658页。7. Jones, C.A. 等，地理位置和支付方类型对加拿大艾伯塔省全省范围内髋膝骨关节炎患者实施*GLA:D*项目的影响。Osteoarthr CarDl Open，2023年。**5**（4）：第100398页。8. Mehta, P. 等，系统性红斑狼疮的胃肠道表现：来自印度多机构起始（*INSPIRE*）队列的数据。 Rheumatology (Oxford)，2025年。**64**（1）：第156-164页。9. Mertens, M.G. 等，多学科多模式疼痛治疗项目治疗成功的预测模型。 Semin ArthriDs Rheum，2025年。**70**：第152592页。10. Palmowski, A. 等，在进入风湿病诊疗前开始使用糖皮质激素，与早期类风湿关节炎老年患者长期使用糖皮质激素相关：*Medicare*与风湿病学信息学有效性系统（*RISE*）数据的联合分析。 Semin ArthriDs Rheum，2024年。**68**：第152535页。11. van Straalen, J.W. 等，甲氨蝶呤治疗与生物制剂初治幼年特发性关节炎患者新发葡萄膜炎发生率降低相关。 RMD Open，2023年。**9**（2）。12. Li, H. 等，*COVID-19*疫苗接种后痛风发作的风险：一项病例交叉研究。 Semin ArthriDs Rheum，2022年。**56**：第152059页。13. Rempenault, C. 等，托珠单抗治疗类风湿关节炎患者与利妥昔单抗或阿巴西普治疗患者相比，憩室炎和胃肠道穿孔的风险。Rheumatology (Oxford)，2022年。**61**（3）：第953-962页。

14. Crossfield, S.S.R. 等，二十年来强直性脊柱炎的发病率、患病率和诊断时间的变化。 RMD Open，2021 年。**7** （ 3 ）。15. Chua, J. 等，利益相关者对卫生服务中骨关节炎干预措施的偏好：一项采用多标准决策分析的横断面研究。 Osteoarthr CarDl Open，2020年。**2**（4）：第100110页。16. Van der Elst, K. 等，五分之一的快速且持续控制的早期类风湿关节炎患者报告治疗*1*年后健康状况不佳。 RMD Open，2020年。**6**（1）。17. Roth, E. 等，胃食管反流病与*EUSTAR*队列中系统性硬化症患者更严重的间质性肺病相关。 Rheumatology (Oxford)，2025年。**64**（Si）：第Si63-si72页。

补充文本-2：
标准化提示词协议。本文件详细说明了用于评估语言模型对 STROBE 报告规范遵循情况的完整三阶段提示序列（包括约束条件、少样本校准示例和目标评估结构）。

#提示-*0* — 评分的角色、边界和逻辑
你将作为一名独立评定者，评估对 STROBE（加强流行病学观察性研究报告）核对清单的遵循情况。
你的任务是严格依据所提供 PDF 文件的内容，评估手稿中是否报告了每一项 STROBE 条目。

评分选项：
- Y（是）：该条目已清晰且明确地报告- N（否）：该条目未报告或仅部分报告- NA（不适用）：该条目从根本上不适用于该研究设计重要规则：

- 本评估仅涉及报告依从性。
- 不要评估方法学质量、正确性或偏倚风险。
- 如果信息不完整、仅为隐含表述、含糊不清或仅部分报告，则必须评分为“N”。

- 仅当该条目确实不适用时才使用“NA”，而不是在信息缺失时使用。- 不要推断或假定手稿中未明确陈述的信息。
#提示-*1*- 校准（*2*篇文章及外部响应）
现提供以下文件：
- “STROBE.pdf”：当前使用的 STROBE清单，作为参考标准。
- “arDcle1.pdf”和“arDcle2.pdf”：两篇观察性研究示例。
- “training_scores.xlsx”：由不属于主要研究评分者的外部人工评估者完成的 arDcle1 和 arDcle2 的 STROBE 依从性评分。
提供这些示例仅用于说明评分规则在实践中的应用方式，以及评定为“是”时所要求的明确程度。
这些内容并非作为金标准答案提供，也不得用于训练、优化或学习，报告阈值校准除外。

请查看STROBE清单，并检查评分规则是如何应用于文章1和文章2的。

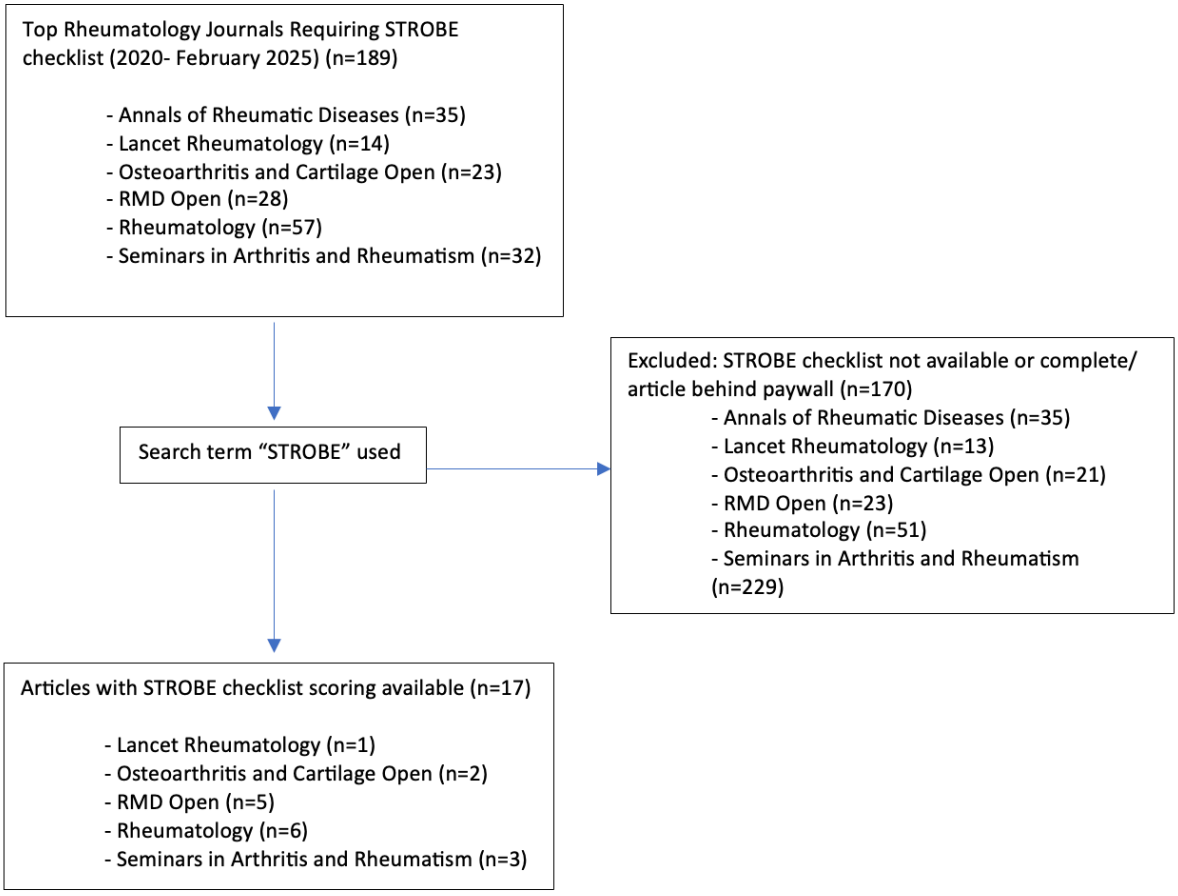
现阶段不要生成任何新的评分。
准备好将相同的评分规则一致地应用于新的手稿后，请予以确认。

#PROMPT-2 - 文章评分
现在请使用之前应用的相同STROBE清单和评分规则，评估所附手稿。

请使用“STROBE.pdf”中提供的 STROBE清单。
输出要求：- 提供一个恰好包含两列的表格：

1) STROBE条目（包括 1a、1b 等子条目） 2) 评分（Y / N / NA）不要提供解释、理由、页码或评论。不要总结或解读结果。不要将这份手稿与之前的示例进行比较。请生成一份完整的 STROBE 依从性表。

补充图-1。研究筛选流程图。



补充图 2：人工审稿人与大语言模型之间的条目特异性评审者间信度。 热图展示了STROBE清单中每个条目和子条目的 Gwet 的 AC1 一致性评分。每个矩阵比较了两种大语言模型（ChatGPT-5.2 和 Gemini 3 Pro）、五名具有不同临床经验的独立人工评定者（初级-1 至高级）以及原研究作者之间的分类报告评估（是/否/不适用）。颜色渐变表示一致性强度，范围从 0.00（一致性较差；浅蓝色）到 1.00（完全一致；深蓝色）。

