L-PRISMA：生成式人工智能（生成式人工智能）时代对 PRISMA 的扩展

Samar Shailendra、Rajan Kadel、Aakanksha Sharma、Islam Mohammad Tahidul、Urvashi Rahul Saxena墨尔本理工学院信息技术与工程学院（*SITE*），澳大利亚墨尔本。s samar@ieee.org, rkadel@mit.edu.au, aasharma@mit.edu.au, mtislam@mit.edu.au, usaxena@mit.edu.au

arXiv:2603.19236v1 [cs.DL] 6 Jan 2026

摘要——**系统综述和荟萃分析优先报告条目（PRISMA）框架为证据综合提供了严谨的基础，但数据提取和文献筛选的手动流程仍然耗时且受限。生成式人工智能（生成式人工智能）的最新进展，尤其是大语言模型（大语言模型），为自动化和扩展这些任务提供了机会，从而提升效率。然而，PRISMA 的核心原则——可复现性、透明度和可审计性——正受到大语言模型固有非确定性以及幻觉和偏见放大风险的挑战。为解决这些局限，本研究将人工主导的综合与生成式人工智能辅助的统计预筛选步骤相结合。人工监督确保科学有效性和透明度，而统计层的确定性则增强了可复现性。所提出的方法系统地完善了 PRISMA 指南，为将生成式人工智能负责任地融入系统综述工作流程提供了路径。**

索引词—**系统综述、PRISMA、生成式人工智能、大语言模型、可复现性、证据综合。**

一、引言

传统上，文献综述在科研生态系统中发挥着至关重要的作用。它们是汇编、评估和传播各学科大量科学研究成果的重要手段。基于此类文献综述撰写的综述论文，在科研群体中日益受到重视。鉴于新研究发表速度迅猛，综述论文发挥着关键作用，能够将复杂且分散的研究发现提炼为连贯、综合且易于获取的叙述，为当前研究和未来探索提供指引。然而，随着人们对综述论文的依赖日益加深，标准化文献综述流程的必要性也逐渐显现，以确保此类论文的问责、可复现性和科学严谨性。

A. PRISMA 背景

2009年，鉴于开展文献综述需要采用标准化、透明且严谨的方法，一种名为系统综述和荟萃分析优先报告条目（*PRISMA*）的系统性文献综述框架应运而生 [1]。作为系统综述基石的 PRISMA框架更新版于 2020 [2] 发布。新版本完善了原有指南，使其内容更加全面、方法更加严谨，从而提高了研究报告的清晰度和可复现性。

尽管采用了系统化方法，PRISMA仍面临一些挑战 [3], [4]。随着已发表研究呈指数级增长，即便精心制定的检索策略也会返回数千条记录，全面筛选因此既耗时又耗费资源。为应对这一难以处理的大量记录，并使整个流程保持在人力可管理的范围内，研究者常常借助关键词、筛选条件、时间范围或词汇相似度工具来筛选结果。然而，这可能导致意外排除不符合所选标准或关键词的相关研究。此外，标题和摘要筛选以及后续全文审查所需的人工工作，会给研究人员带来沉重的认知负担，并增加出现不一致或人为错误的可能性。因此，尽管PRISMA严谨且结构清晰，系统综述仍受限于检索范围不足、人工筛选不够完善，以及实际上不可能详尽审查海量文献等问题。

近年来，自然语言处理（NLP）取得了巨大进展，使计算机能够更好地理解、诠释和分析人类语言。这些进步，尤其是 Transformer架构 [5]，显著提升了机器在具备上下文感知能力的情况下处理大量非结构化文本的能力。在这些技术进步的推动下，生成式人工智能（GenAI）为理解语言以及生成连贯、切合语境且近似人类表达的文本开辟了新的可能性。

大语言模型（LLM），例如生成式预训练 Transformer（GPT） [6], [7],Claude [8], 和 Gemini [9], ，以及 BERT 等基于Transformer的工具，结合这些能力，大规模地理解和生成信息。这些模型能够提取关键见解、总结研究结果，并适应特定领域的需求，展现出在研究密集型学科中的变革潜力。研究人员可以利用大语言模型自动化并加速繁重的筛选和数据提取流程，从而拓宽系统综述的范围。然而，这些强大却不透明的工具在整合过程中带来了一系列新的根本性挑战，直接与 PRISMA 的核心准则——透明度、可复现性、伦理使用和可审计性[4], [10]——相冲突。

B. 生成式人工智能时代 *PRISMA* 框架面临的问题

生成式人工智能能够成功应对规模和人工工作量方面的挑战，而这些正是限制 PRISMA框架的因素 [3], [4]。首要挑战在于可复现性。传统的 PRISMA检索策略基于特定关键词和数据库查询，旨在准确记录并供其他研究人员复现。然而，生成式人工智能驱动的发现往往具有非确定性。结果可能因所用模型及其版本、模型产生幻觉的情况，以及提示方式的细微差异而有所不同。这种“黑箱”特性使得人工智能辅助检索过程极难报告，也无法保证其他研究人员能够得出相同结果。为解决这一问题，我们提出了一种生成式人工智能辅助的统计方法，即使用统计方法进行预筛选，以确保结果具有确定性和可复现性，同时避免研究人员需要深入掌握统计学知识。

筛选和数据提取的自动化引发了人们对可靠性和科学严谨性的担忧。大语言模型可能会误解细微的科学论证，无法区分高质量研究与有缺陷的研究，或生成“幻觉”——看似可信、实则完全捏造的引文和摘要 [22], [23]。这使研究者的重心从人工筛选转向同样艰巨的任务：细致核验人工智能的输出。作者身份以及对最终报告结果所负知识责任的问题，因偏倚风险而变得更加复杂。因此，我们建议按照传统的人工方式审查最相关的文献，同时有策略地纳入基于生成式人工智能的检索，以识别可能被遗漏的研究，从而在不牺牲严谨性的前提下尽可能提高全面性。

Gundersen等人 [11] 的研究有力地证明了开放科学流程在人工智能领域的有效性，表明可复现性既是科学层面的前提，也是技术层面的前提，对于维持问责并促进累积性研究至关重要。他们的初步发现与生成式人工智能驱动的文献综述所面临的困难一致：大语言模型的非确定性会损害可复现性。Haibe-Kains等人 [12] 提出的研究为理解非确定性人工智能方法（大语言模型）在系统综述和类PRISMA工作流程中的局限性奠定了基础。这凸显了开放且一致的流程的重要性，尤其是在将人工智能纳入PRISMA等证据综合框架时。此外，Schwartz等人 [13] 提出了一种系统的方法，用于识别和控制人工智能中的偏见，并提供了减少自动化筛选和数据提取相关风险的实用策略。他们的观点与Lloyd [14],的看法一致；Lloyd提醒说，人工智能系统可能通过反馈回路强化既有偏见，从而扭曲研究的广度及其解读。总的来说，这些结果表明，尽管生成式人工智能为提升系统综述的规模化能力带来了机遇，但在应用时仍须谨慎，并采取措施保障可复现性、透明度和偏见减少，以维护科学诚信。此外，为确保人工智能增强的系统综述在设计、收集和解读各阶段都具备透明度、有效性和可复现性，Malik和Terzidis [28]提出了一种混合方法，以兼顾计算效率和认识论严谨性。

近年来，将人工智能融入系统综述工作流程的进展，已开始解决证据综合中与规模扩展相关的问题。众多研究表明，生成式人工智能技术能够自动执行数据提取和筛选，降低人工劳动成本，并扩大可检索文献的范围 [15]。然而，大语言模型驱动的流程会产生非确定性输出，使透明度和独立验证更加困难，因此可重复性问题依然存在 [24]。与此同时，关于混合框架的研究强调，将确定性统计技术与基于人工智能的自动化相结合，对于解决可重复性问题并确保方法学稳健性至关重要 [25]。此外，关于人工智能辅助审查流程的研究强调了偏差缓解和检测技术的重要性，并提醒人们，盲目依赖机器生成的输出可能会使证据综合中的持续性错误长期存在 [26]。此外，为使人工智能增强程序与 PRISMA 等成熟规范保持一致，并维护科学严谨性和可信度，新兴系统还提供公开报告要求和规范化的审计追踪记录 [27]。这些研究成果既支持生成式人工智能加快系统综述进程的潜力，也强调了建立保障措施的迫切需要，以维持可复现性、透明度和偏倚控制。表 I 总结了生成式人工智能时代应用 PRISMA 框架所面临的问题。

C. 动机与贡献

基于上一节讨论的 PRISMA 框架的局限性，我们提议对 PRISMA 框架进行完善，将基于生成式人工智能的工具纳入其中，同时不损害 PRISMA 已确立的科学严谨性和透明度。本文为利用这些强大的技术提升系统综述的质量提供了清晰的路径。本文的主要贡献如下：

增强的 ***PRISMA*** 框架：修订后的 PRISMA清单和更新后的流程图，明确纳入了生成式人工智能应用的各个阶段。

系统化整合生成式人工智能：一套用于系统综述的统计方法（借助生成式人工智能），并提供报告指南以确保输出验证。这将帮助读者复现系统检索结果。

一种混合综述方法： 一种由生成式人工智能增强的综述方法，以确保传统人工主导的综述作为核心组成部分的可靠性，同时系统地将生成式人工智能用作辅助发现工具，以扩大检索范围并识别

表 I 在生成式人工智能时代应用 PRISMA框架面临的问题。

Area	问题	说明
检索策略透明度	人工智能辅助搜索 may not be 可复现性 [11], [12]	人工智能生成的检索策略（例如，来自 ChatGPT 的策略）可能无法提供完整的透明度或一致性，因此难以复现系统性检索。
选择偏倚	生成式人工智能可能引入或加剧偏见 [13], [14]	大语言模型基于现有数据进行训练，可能会优先考虑某些来源或观点，从而可能使文献筛选产生偏差。
筛选 and 资格判定	过度依赖 on 生成式人工智能 for 纳入/排除决策	将决策权委托给人工智能工具可能会损害人类判断，并削弱语境理解，尤其是在纳入标准较为微妙复杂的情况下。
数据提取	人工智能自动化可能会误解细微的《数据》 [15]	人工智能可以帮助提取结构化信息，但它可能难以处理依赖具体情境或特定学科领域的数据解读。
批判性评价	缺乏严谨的质量评估 人工智能 [16]	人工智能缺乏细致评估研究质量所需的批判性思维，可能导致纳入质量低下或有偏倚的研究。
更新与持续更新的综述	自动更新可能会绕过人工监督 [17]	人工智能可以协助开展“动态系统综述”，但如果缺乏严谨的验证，自动更新可能会纳入未经验证或质量低劣的来源。
抄袭 & 重复风险	人工智能生成的内容可能会反映训练数据 [18]	大语言模型生成的摘要或解读可能会无意中复现或近乎照搬现有文献，却未给予适当引用。
PRISMA Flow 图表生成	自动化工具可能错误呈现工作流程 [19]	某些基于生成式人工智能的工具可能生成不完整或过度简化的图表，无法准确反映综述过程。
伦理 and 作者身份	人工智能在作者身份中的作用不明确 [20], [21]	对于如何在系统综述中注明生成式人工智能的贡献，缺乏明确说明带来伦理和学术困境。
对人工智能的过度自信输出	幻觉或不准确的引文 [22], [23]	生成式人工智能可能会生成看似可信但实为捏造的参考文献，或错误归属研究发现，从而导致系统综述中出现错误信息。
人工智能增强系统综述	人工筛选和数据提取既耗时又难以扩展 [15]	在确保计算效率的同时，保持认识论标准即透明度、可复现性和严谨性，从而避免过度依赖不透明模型，同时利用其优势来减少工作量。
人工智能工作流中的可复现性工作流	大语言模型的输出具有非确定性，因此结果复现颇具挑战性 [24]。	指出生成式人工智能带来的可复现性差距，并强调以下需求：确定性方法或报告规范，以确保透明度。
混合框架	纯人工智能驱动的流程存在出错风险，且缺乏稳健性 [25]	提出将统计预筛选与人工智能方法相结合，以平衡效率与可复现性，确保方法学严谨性。
自动化 of 系统综述	手动文献综述 are 耗时，耗费大量资源，and 常因.....的快速增长而受阻出版物 [26]	人工智能驱动的方法可以自动完成系统综述的关键阶段，例如文献检索、筛选和数据提取，从而提升可扩展性同时保持方法学严谨性。
透明度 and 可审计性	Lack 的可追溯性 in 人工智能增强 PRISMA 流程 [27]	倡导建立结构化审计追踪、提供透明文档并遵循既定标准（例如 PRISMA），以维护科学诚信。

常规关键词策略可能遗漏的文献。

本文的结构如下：第二节简要回顾相关文献。第三节详细介绍大语言模型增强的系统综述和荟萃分析首选报告条目（L-PRISMA）框架，第四节则通过一个用例示例进行分析。最后，第五节提出总结性意见和未来方向。

II. 文献综述

本节介绍系统综述领域的近期研究、针对文献综述提出的各类工具，以及用于系统综述的不同工具。

Toth等人 [29] 梳理并评估了可用于支持系统综述的自动化工具，并按照PRISMA工作流程阶段（检索、筛选、数据提取、综合和报告）对其进行分类，同时详细说明了如何利用这些工具简化证据综合。然而，这项工作并未充分利用现代自然语言处理工具和大语言模型的真正潜力，以拓展文献综述的广度和深度。在 [30], 中，作者提出了一种半自动化的文献综述方法，但没有概述任何用于筛选应接受人工审查的研究的系统性标准。缺乏明确的筛选框架会增加无意中排除相关研究或纳入无关研究的风险。因此，有必要建立一种有原则的机制，以指导纳入文献进行系统综述和人工审查。

Kolaski等人 [31], 将 PRISMA 与其他现有标准（AMSTAR-2、ROBIS 和 GRADE）进行评估和比较，并将其总结为一份简明指南，介绍各工具的功能，同时列出有助于改进系统综述实施和报告的最佳实践资源。不过需要注意的是，与 PRISMA 不同，系统综述测量工具（AMSTAR-2） [32] 主要适用于医学和医疗保健相关领域，而系统综述偏倚风险评估工具（ROBIS） [33] 和建议评估、制定与评价分级（GRADE） [34] 分别用于处理系统综述的偏倚风险和质量。此外，他们的工作缺少针对系统综述的可操作或经验证的方案/框架，尤其是能够适应生成式人工智能这一新兴时代的方案/框架。

在 [35], 作者介绍了系统综述设施（SyRF），这是一个专为支持临床前系统综述而设计的网络平台，提供筛选、数据管理和协作工作流程等功能；然而，这项工作仅限于主要属于

医学领域的临床前研究，无法推广到其他领域，因此适用范围较窄。文献中还提出了许多文献综述工具（例如 Rayyan [36], Covidence [37],RevMan [38], ASReview [39]）。由于该研究仅涉及两篇系统性文献综述、特异度值偏低至中等，且依赖训练集的质量，其发现可能不适用于其他工具或综述。该分类器用于检测随机对照试验（RCTs） [37]，但它识别为可能属于随机对照试验的论文中，有相当一部分实际上并非随机对照试验。用户仍须手动检索大量无关文献，才能找到真正的随机对照试验。由于该论文采用的统计工具有限、用户友好性不足，且缺乏图形调整选项，其发现仅适用于消费者 [38]。更复杂的 ASReview 模型（如 Doc2Vec 和 SBERT）对计算资源的需求，可能导致处理时间延长，尤其是在处理大型数据集时 [39]。尽管这些工具目前会使用人工智能来优化和总结文献，但仍缺少 PRISMA框架所提供的系统性方法。此外，人工智能训练的局限性、统计工具的复杂性，以及面对规模超出预期的数据集时处理时间较长，都会使高效获得结果变得困难。最近，Teo [40] 提出了一个针对特定领域微调、基于大语言模型的 PRISMA 扩展。然而，这需要使用特定领域的文献对大语言模型进行微调，通常耗费大量资源和时间，并需要具备专门的模型微调技能。这削弱了使用该模型的便利性，使得这种方法只有深入了解大语言模型的研究人员才能使用，也违背了该框架原本面向所有人的简便性和可及性。还有一些大语言模型（如 Perplexity [41]）声称专为研究设计，并通过与实际参考资料交叉核对来优化模型，避免产生幻觉或提供非事实内容。

对当前文献的综述表明，传统的文献综述方法往往缺乏研究筛选标准，可能导致必要信息被排除。尽管半自动化方法有所改进，但仍需要进行全面的人工

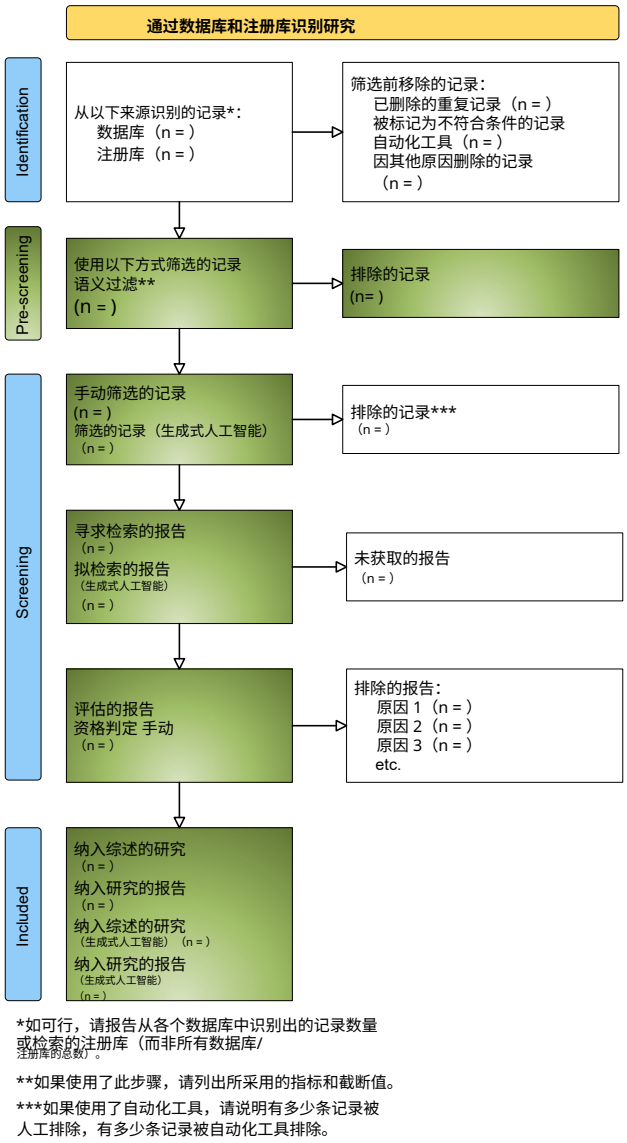


图1. L-PRISMA流程图。

检查无关内容，而许多平台具有特定领域的局限性，例如医学领域，这限制了它们的广泛应用。此外，在当前人工智能时代，缺少开展彻底且可重复评估所需的结构化、全面框架。因此，需要一个更全面的框架，以提供系统、严谨的方法来指导文献的系统纳入，确保综述全面、客观且可复现，从而适应生成式人工智能等新兴和不断发展的技术。

III. L-PRISMA框架

PRISMA一直是系统性文献综述中最成功的框架。基于第二节中的广泛文献综述以及我们自身的经验 [42], [43], , 我们发现该框架存在一些局限性，如果能够解决这些局限性，PRISMA框架将显著提升系统性文献综述的质量和深度。因此，PRISMA有必要正式将生成式人工智能纳入其中，作为其不可或缺的一部分，并制定使用生成式人工智能开展系统性文献综述的明确指南。

本文提出了 L-PRISMA，这是生成式人工智能时代对 PRISMA 框架的扩展，旨在用于系统性文献综述。我们的框架利用语义过滤、基于 Transformer 的模型和基于大语言模型的工具来生成文献摘要。此外，它还提供统计指南，以进一步优化并筛除基于工具的搜索结果。与文献中的其他方案不同，该框架对资源的要求较低，非 IT 专业研究人员也能轻松使用，这符合 PRISMA 的核心理念：轻量且人人可用。还需指出的是

由于大语言模型往往会产生幻觉 [22], [23] 并自信地生成错误结果，因此在整个过程中需要谨慎行事，由人工审核所有结果。

L-PRISMA框架（图1）提出了对 PRISMA 的以下更新：（i）新增预筛选阶段，用于语义过滤；（ii）更新筛选阶段，以报告生成式人工智能记录；（iii）更新纳入阶段，以报告生成式人工智能报告和研究。在预筛选阶段，使用语义过滤对检索到的记录进行筛选，同时利用生成式人工智能总结该阶段归类的部分记录。下文小节将讨论这些阶段的具体细节。

A. 预筛选阶段

与 PRISMA 建议在筛选阶段使用工具过滤记录不同，L-PRISMA 将语义过滤作为预筛选阶段加入其中；此后可按照 PRISMA 的描述，使用词汇相似度工具进行筛选。在预筛选期间，根据搜索标准检索到的论文数量可能很大。需要注意的是，通过限制关键词或正则表达式来缩小搜索条件，可能会严重限制搜索范围，并遗漏一些重要成果。在此阶段，我们使用语义相似度分数筛选出相关论文。为计算该相似度分数，作者会定义一段描述其综述意图的陈述，并据此计算语义相似度分数。随后绘制所得分数的直方图。我们可以将相似度分布近似为两个（或更多个）相互重叠的组的组合。在不失一般性的情况下，该分布可表示为两个（或更多个）分布的混合：一个对应高度相关的文章，另一个对应相关性较弱（或不相关）的文章，如基于相似度搜索 [44]–[46] 的公式（1）所示。

$$\begin{aligned} p(s) &= \pi_H\, p_H(s) + \pi_L\, p_L(s), \quad s \in [0,1] \\ \pi_H &+ \pi_L = 1 \end{aligned} \tag{1}$$

其中，
$p(s)$表示总体相似度分数分布，
p_H 表示高相关性分数分布，
p_L 表示低相关性分数分布，以及
$\pi_H,\ \pi_L \geq 0$ 为混合权重。
随后，可以使用四分位点或任何其他统计边界规则来确定决策边界。

此阶段将记录筛选阶段中被排除的记录，以及标记为需要人工筛选的记录。我们建议使用大语言模型解决这一问题。作者不一定需要了解任何统计工具。这样，用户便可以在充分知情的情况下决定纳入多少篇论文进行人工审查，纳入多少篇进行生成式人工智能审阅，或将其排除。下一节将通过一个用例示例来展示这种方法。

B. 筛选阶段

L-PRISMA 的这一阶段与标准 PRISMA框架相同，不同之处在于，除报告待筛选的记录数量外，还需报告待手动筛选的记录数量，以及使用生成式人工智能进行筛选和总结的记录数量。还应报告基于人工筛选而排除的记录数量。

C. 纳入阶段

传统上，PRISMA 默认综述中纳入的所有研究均由人类开展。人工审查对于文献综述至关重要，也是其基石，但人类能够审查的研究数量存在实际限制。这也可能导致一些相关性很高的研究被遗漏在综述之外。为缓解这一问题，我们建议将研究分为两部分：相关性较高的报告（例如相似度分数较高的报告）应由人工总结；而相关性较低或较弱的报告，则可在适当提示词的引导下由生成式人工智能总结，使总结结果与整体综述结构保持一致。这样可以确保系统综述不会舍弃或遗漏文献中的重要方面。还需注意，生成式人工智能的结果并非完美，因此始终应由人工监督和检查其一致性。

IV. 用例示例

本节介绍一个实际用例场景的结果，并说明 L-PRISMA 的附加阶段如何帮助系统综述流程。

A. 第一阶段：识别

我们在两个数据库——IEEE 和 ACM——中进行了检索。所使用的检索式以及返回的记录数量列于表 II。我们的初衷是了解生成式人工智能工具如何在教育领域用于文本相似度研究。我们使用了两类检索式：一类在检索式中包含“Education”一词，另一类则不限定具体的教育领域。显然，如果将检索限定为教育领域的研究，检索结果数量会更易于处理，也与教育领域高度相关；然而，很容易看出，这样会遗漏与文本相似度和生成式人工智能相关的重要文献。为了使结果数量保持在人力可处理的范围内，研究者可能会采用限制性更强的检索方式，但这也有失去大量相关重要文献的风险。不过，L-PRISMA 不受这一问题的限制，并建议纳入所有检索到的记录。同时，我们也承认，范围更广的检索可能会包含一些与原始文献综述主题无关的研究。为缓解这一问题，我们将在下一节讨论并提出预筛选阶段，利用生成式人工智能的能力，以统计方式剔除噪声文献。

表 II 不同数据库的检索式及结果数量（检索日期：2025年8月1日）

数据库	检索式	检索范围	记录数
IEEE	(((“语义” OR “相似性”):摘要 OR (“语义” OR “相似性”):“文档标题”) AND ((“自然语言处理” OR 自然语言处理 OR “生成式人工智能” OR 生成式人工智能 OR “生成式人工智能”):摘要 OR (“自然语言处理” OR 自然语言处理 OR “生成式人工智能” 或 生成式人工智能 或 “生成式人工智能”）：“文档标题”）且（（“教育”）：摘要 或（“教育”）：“文档标题”））	With 教育 领域 约束	24
ACM	((标题:(语义 OR 相似性) OR 摘要:(语义 OR 相似性)) AND (标题:(“自然语言处理” OR 自然语言处理 OR “生成式人工智能” OR “生成式人工智能” OR “生成式人工智能”) OR 摘要:(“自然语言处理” OR 自然语言处理 OR “生成式人工智能” OR “生成式人工智能” OR “生成式人工智能”)) AND (标题:(教育) OR 摘要:(教育)))	With 教育 领域 限制条件	48
IEEE	(((“语义” OR “相似性”):摘要 OR (“语义” OR “相似性”):“文档标题”) AND ((“自然语言处理” OR 自然语言处理 OR “生成式人工智能” OR 生成式人工智能 OR “生成式人工智能”):摘要 OR (“自然语言处理” OR 自然语言处理 OR “生成式人工智能” OR 生成式人工智能 OR “生成式人工智能”):“文档标题”))	不含 教育 领域 限制	362
ACM	((标题:(语义 OR 相似性) OR 摘要:(语义 OR 相似性)) AND (标题: (“自然语言处理” OR 自然语言处理 OR “生成式人工智能” OR “生成式人工智能” OR “生成式人工智能”) OR 摘要: (“自然语言处理” OR 自然语言处理 OR “生成式人工智能” OR “生成式人工智能” OR “生成式人工智能”)))	无 教育 领域 限制	869

B. 第二阶段：预筛选

这是 L-PRISMA 引入的新阶段。如前所述，我们希望在此阶段筛选出与文献综述相关的记录。该阶段使用记录的标题和摘要作为识别依据。首要任务是定义一句能够最准确描述文献综述意图的陈述。为此，我们明确了研究意图，并请不同的生成式人工智能工具（ChatGPT、Gemini 和 Claude）对陈述进行润色。最终，我们形成了如下综述陈述：“本研究探讨衡量学生生成的回答与参考答案之间文本和语义相似度的方法，重点关注利用自然语言处理技术进行自动评分和教育评估的应用。”

请注意，使用生成式人工智能构思此类陈述并非强制要求，但可能会有所帮助，尤其是对于英语非母语者。还值得注意的是，这条检索陈述包含了研究者感兴趣的完整语境，因为这有助于人工智能工具更好地理解他们的意图。

其次，我们决定使用句子双向转换器表示（S-BERT）模型对句子对进行微调，然后使用余弦相似度来查找调查陈述与记录 [47] 之间的语义相似度。S-BERT 将句子转换为向量，从而能够高效地衡量语义相似度，因此非常适用于需要理解句子或文档间含义的自然语言处理任务。此外，这是迄今为止用于捕捉句子相似性的最佳生成式人工智能模型之一，但研究者也可以使用其他相似性模型和衡量矩阵。

这些记录的相似度分数的概率密度函数（PDF）绘制在图 2 中。从图中可以清楚地看出，搜索结果的相似度分数遵循高斯混合模型（GMM）。随后，根据公式（1），很容易推断出

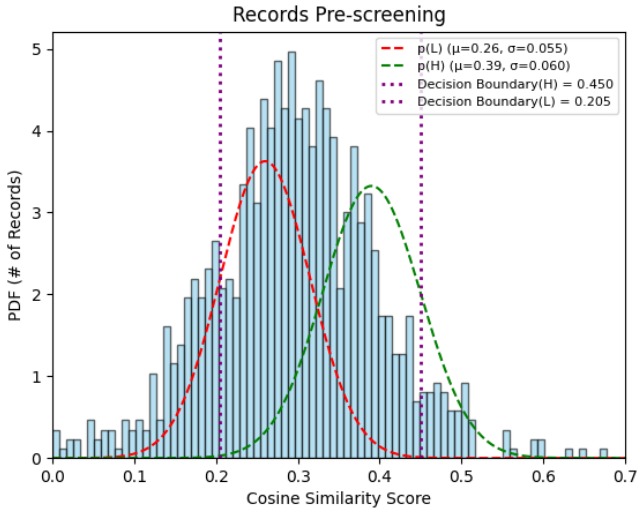


图 2. 记录数量的概率密度函数（概率密度函数）。

该高斯混合模型可以视为公式（2）所述的两个高斯分布的组合 [46]。

$$p(s) = \sum_{k=H,L} \pi_k \mathcal{N}(s_i \mid \mu_k, \sigma_k^2), \tag{2}$$

其中
s_i 是第 i^{th} 个数据点（即余弦相似度分数），
π_k 是混合权重，其中 $\sum_j \pi_j = 1$,
μ_k, σ_k^2 分别是成分 k的均值和方差，
$\mathcal{N}(x_i \mid \mu_k, \sigma_k^2)$ 是在成分 k下对 x_i 求得的高斯密度。

我们使用 Python GMM 库获得了这两个高斯分布的分布。由于二者均服从高斯分布（$p_k(s) \sim \mathcal{N}(\mu_k, \sigma_k^2)$），因此我们使用 $2\sigma_k$ bounds 获得了决策边界。结果绘制在图 2 中。在此阶段，低于较低截断点的记录将被移除（n=182）。

高于上限截断值的记录（n=60）是最相关的记录，需要在筛选阶段手动筛选；而位于两个决策边界之间的记录则可视为已筛选，并由生成式人工智能进行总结（n=989）。还需指出，这些数字仅为指导值，研究人员可结合其领域专业知识进一步调整。

C. 阶段三：筛选

本阶段遵循标准的 PRISMA筛选方法，并进行了以下修改。通过预筛选阶段识别出的记录将接受资格评估筛选；语义相似度分数较高（高于预筛选阶段确定的统计阈值）的记录（n=60）由人工审阅者进行传统的人工筛选，而语义相似度分数较低的记录（n=989）则使用与资格标准相符的结构化提示词进行生成式人工智能辅助筛选。在全文检索阶段，将分别为人工评估（高相关性记录）和生成式人工智能辅助评估（低相关性记录）获取报告。任何无法检索到的报告都应予以记录。随后，对相似度分数较高的报告进行人工评估，并记录排除的报告及其排除理由。务必记录整个过程中所使用的具体大语言模型和提示词。

D. 阶段四：纳入

本阶段报告纳入研究以及这些研究中的报告数量。此处唯一的变化是，研究既要报告由人类进行的人工审查结果，也要报告生成式人工智能审查结果。生成式人工智能审查由大语言模型执行，并且必须提供所使用大语言模型的详细信息。

V. 结论与未来方向

总之，本研究介绍了 L-PRISMA，这是对 PRISMA框架的增强型改编，纳入了生成式人工智能在系统性文献综述方面的最新进展，从而提高了方法效率。为应对大语言模型固有的非确定性，本研究进一步将人工主导的综合与生成式人工智能辅助的统计预筛选流程相结合。为详细说明框架的各个步骤，研究还提供了一个用例，并进行了相应的统计分析。未来研究应将该流程应用于不同领域和文献，以进一步完善该流程。此外，还需开展进一步研究，以适应生成式人工智能不断涌现的能力，并开发领域特定方法，从而提高 L-PRISMA框架的可靠性并扩大其适用范围。

参考文献

[1] D. Moher、A. Liberati、J. Tetzlaff、D. G. Altman 和 T. P. Group，“系统综述和荟萃分析优先报告条目：PRISMA声明”，*BMJ*，第 339 卷，页码 b2535，2009年。[在线]。网址：http://www.bmj.com/content/339/bmj.b2535.abstract

[2] M. J. Page、J. E. McKenzie、P. M. Bossuyt、I. Boutron、T. C. Hoffmann、C. D. Mulrow、L. Shamseer、J. M. Tetzlaff、E. A. Akl、S. E. Brennan 等，“PRISMA 2020声明：系统综述报告指南的更新”，BMJ，第372卷，2021年。[3] J. P. Ioannidis，“冗余、误导性和存在利益冲突的系统综述与荟萃分析的大量产出”，《米尔班克季刊》，第94卷，第3期，第485–514页，2016年。[4] D. Kr´ol 和 M. Kutrzy´nski，“关于在系统综述中负责任地使用自动化”，载于《智能信息与数据库系统的最新挑战》。新加坡施普林格·自然，2025年，第308–321页。[5] A. Vaswani、N. Shazeer、N. Parmar、J. Uszkoreit、L. Jones、A. N. Gomez、L. Kaiser 和 I. Polosukhin，“《Attention Is All You Need》”，载于第31届神经信息处理系统国际会议论文集，NIPS’17系列。美国纽约州雷德胡克：Curran Associates Inc.，2017年，第6000–6010页。[6] A. Radford 和 K. Narasimhan，“通过生成式预训练改进语言理解”，2018年。 [在线]。可访问：https://api. semanticscholar.org/CorpusID: 49313245

[7] G. Yenduri、M. Ramalingam、G. C. Selvi、Y. Supriya、G. Srivastava、P. K. R. Maddikunta、G. D. Raj、R. H. Jhaveri、B. Prabadevi、W. Wang 等，“GPT（生成式预训练 Transformer）——关于赋能技术、潜在应用、新兴挑战和未来方向的综合综述”，IEEE Access，第12卷，第54 608–54 649页，2024年。[8] Y. Bai、S. Kadavath、S. Kundu、A. Askell、J. Kernion、A. Jones、A. Chen、A. Goldie、A. Mirhoseini、C. McKinnon 等，“宪法式人工智能：来自人工智能反馈的无害性”，arXiv预印本 arXiv:2212 08073，2022年。。[9] G. T. Google，“Gemini：高能力多模态模型系列”，2025 [ ] 2312 11805 。在线。可访问：https://arxiv.org/abs/。

[10] D. Eacersall、L. Pretorius、I. Smirnov、E. Spray、S. Illingworth、R. Chugh、S. Strydom、D. Stratton-Maher、J. Simmons、I. Jennings 等，“应对生成式人工智能增强研究中的伦理挑战：负责任地使用生成式人工智能的伦理框架”，arXiv预印本 arXiv:2501.09021，2024年。[11] O. E. Gundersen、O. Cappelen、M. Møln˚a 和 N. G. Nilsen，“开放科学在人工智能领域的非凡有效性：一项重复研究”，载于人工智能促进协会人工智能会议论文集，第39卷，第25期，2025年，第26 211–26 219页。[12] B. Haibe-Kains、G. A. Adam、A. Hosny、F. Khodakarami、M. A. Q. C. M. S. B. of Directors Shraddha Thakkar 35 Kusko Rebecca 36 Sansone Susanna-Assunta 37 Tong Weida 35 Wolfinger Russ D. 38 Mason Christopher E. 39 Jones Wendell 40 Dopazo Joaquin 41 Furlanello Cesare 42、L. Waldron、B. Wang、C. McIntosh、A. Goldenberg、A. Kundaje 等，“人工智能中的透明度与可复现性”，《自然》，第586卷，第7829期，第E14–E16页，2020年。[13] R. Schwartz、R. Schwartz、A. Vassilev、K. Greene、L. Perine、A. Burt 和 P. Hall，《人工智能中的偏见识别与管理标准》，美国商务部、美国国家标准与技术研究院……，2022年，第3卷。[14] K. Lloyd，“人工智能系统中的偏见放大”，arXiv预印本 arXiv:1809.07842，2018年。[15] A. Adel 和 N. Alani，“生成式人工智能能否可靠地综合文献？探索ChatGPT中的幻觉问题”，《人工智能与社会》，2025年。[16] J. Zybaczynska、M. Norris、S. Modi、J. Brennan、P. Jhaveri、T. J. Craig 和 T. Al-Shaikhly，“人工智能生成的科学文献：批判性评价”，《过敏与临床免疫学杂志：临床实践》，第12卷，第1期，第106–110页，2024年。[在线]。可访问：https://www.sciencedirect.com/science/article/pii/ S2213219823011261

[17] F. Bola˜nos、A. Salatino、F. Osborne 和 E. Motta，“用于文献综述的人工智能：机遇与挑战”，《人工智能评论》，第57卷，第10期，第259页，2024年8月。[ ] 10 1007 10462 024 10902 3 在线。可访问：https://doi.org/。/s - - - J. P. Wahle、T. Ruas、N. Meuschke 和 B. Gipp，“神经语言模型是优秀的剽窃者吗？神经释义检测的基准测试”，载于2021年ACM/IEEE数字图书馆联合会议（JCDL），2021年，第226–229页。[19] D. Kr´ol 和 M. Kutrzy´nski，“关于在系统综述中负责任地使用自动化”，载于《智能信息与数据库系统的最新挑战》，N. T. Nguyen、T. Matsuo、F. L. Gaol、Y. Manolopoulos、H. Fujita、T.-P. Hong 和 K. Wojtkiewicz 编，新加坡：新加坡施普林格·自然，2025年，第308–321页。

[20] G. Andrade-Hidalgo、P. Mio-Cango 和 O. Iparraguirre-Villanueva，“探索人工智能对研究伦理的影响——系统综述”，《学术伦理杂志》，第23卷，第3期，第1053–1070页，2025年9月。 [在线]。可访问：https: //doi.org/ 10 1007 10805 024 09579 8 。/s - - - S. Knight，“通过受[21] PRISMA-ETHICS启发的指南范围综述，理解人工智能伦理指南制定中的证据使用”，《计算机与教育开放》，第100281页，2025年。[在线]。可访问：https://www.sciencedirect.com/science/article/pii/ S2666557325000400[22] Z. Ji、N. Lee、R. Frieske、T. Yu、D. Su、Y. Xu、E. Ishii、Y. J. Bang、A. Madotto 和 P. Fung，“自然语言生成中的幻觉调查”，《ACM计算评论》，第55卷，第12期，2023年3月。 [在线]。可访问：https://doi.org/ 10 1145 3571730 。/ V. Rawte、A. Sheth 和 A. [23] Das，“大型基础模型中的幻觉调查”，2023年。 [在线]。可访问：https: // arxiv.org/abs/ 2309 05922 。F. Bolanos、A. Salatino、F. [24] Osborne 和 E. Motta，“用于文献综述的人工智能：机遇与挑战”，《人工智能评论》，第57卷，第10期，第259页，2024年。[25] L. Ge、R. Agrawal、M. Singer、P. Kannapiran、J. A. De Castro Molina、K. L. Teow、C. W. Yap 和 J. A. Abisheganaden，“利用人工智能增强健康研究中的系统综述：先进工具与挑战”，《系统综述》，第13卷，第1期，第269页，2024年。[26] J. de la Torre-L´opez、A. Ram´ırez 和 J. R. Romero，“利用人工智能自动化科学文献系统综述”，《计算》，2023年。[27] L. Li、A. Mathrani 和 T. Susnjak，“转变证据综合：人工智能时代自动化荟萃分析演进的系统综述”，arXiv预印本 arXiv:2504.20113，2025年。[28] A. Malik 和 O. Terzidis，“人工智能增强系统综述的混合框架：在计算效率与认识论严谨性之间取得平衡”，《商业经济学杂志》，2025年。[29] B. T´oth、L. Berek、L. Gul´acsi、M. P´entek 和 Z. Zrubka，“生物医学文献系统综述的自动化：PubMed收录研究的范围综述”，《系统综述》，第13卷，第1期，第174页，2024年。[30] L. Schmidt、A. N. Finnerty Mutlu、R. Elmore、B. K. Olorisade、J. Thomas 和 J. P. T. Higgins，“系统综述数据提取方法的（半）自动化：一项持续更新的系统综述的更新”，F1000Research，第10卷，第401页，2021年。 [在线]。可访问：https://doi.org/ 10 12688 1000 51117 3 。/f researc [31] h. 。K. Kolaski、L. R. Logan 和 J. P. A. Ioannidis，“系统综述的最佳工具与实践指南”，《系统综述》，第12卷，第1期，第96页，2023年。[32] B. J. Shea、B. C. Reeves、G. Wells、M. Thuku、C. Hamel、J. Moran、D. Moher、P. Tugwell、V. Welch、E. Kristjansson 等，“AMSTAR 2：用于纳入随机或非随机研究医疗干预措施（或两者）的系统综述的批判性评价工具”，BMJ，第358卷，2017年。[33] P. Whiting、J. Savovi´c、J. P. T. Higgins、D. M. Caldwell、B. C. Reeves、B. Shea、P. Davies、J. Kleijnen 和 R. Churchill，“ROBIS：一种用于评估系统综述偏倚风险的新工具”，《学术伦理杂志》《临床流行病学》，第 69 卷，页 225–234，2016年。 [在线]。获取地址：https://doi.org/ 10 1016 2015 06 005 。/ [34] j.jclinepi. . . G. Guyatt、T. Agoritsas、R. Brignardello-Petersen、R. A. Mustafa、J. Rylance、F. Foroutan、M. Prasad、A. Agarwal、H. De Beer 等，“Core GRADE 1：Core GRADE 方法概述”，《英国医学杂志》，第 389 卷，2025年。[35] Z. Bahor、J. Liao、G. Currie、C. Ayder、M. Macleod、S. K. McCann、A. Bannach-Brown、K. Wever、N. Soli man、Q. Wang 等，“在线系统综述平台的开发与应用：CAMARADES 系统综述设施（SyRF）的早期发展”，《英国医学杂志开放科学》，第 5 卷，第 1 期，文章编号 e100103，2021年。[36] J. Ng、N. Szydlowski、M. Gill、N. Fusco 和 K. Ruiz，“MSR72：对 Rayyan 人工智能工具在系统文献综述筛选中的评估”，《健康价值》，第 27 卷，273，2024 6，第 S 页。[37] K. Cowie、A. Rahmatullah、N. Hardy、K. Holub 和 K. Kallmes，“医学系统性文献综述的基于网络的软件工具：系统检索与功能分析”，《*JMIR* 医学信息学》，第 10 卷，第 5 期，文章编号 e33219，2022年。[38] J. Wang 和 M. Leeflang，“诊断准确性荟萃分析推荐软件/软件包”，《实验室与精准医学杂志》，第 4 卷，期号未知，页 1–13，2019年。[39] Y. Quan、T. Tytko 和 B. Hui，“利用 ASReview 筛选用于 SLA 元研究的原始研究：分步教程”，《应用语言学研究方法》，第 3 卷，第 1 期，文章编号 100101，2024年。[40] T. Susnjak，“PRISMA-DFLLM：面向使用领域特定微调大语言模型开展系统性文献综述的 PRISMA 扩展”，arXiv，2023年。 [在线]。获取地址：https: //arxiv.org/abs/2306.14905[41] Perplexity AI，“Perplexity AI”，Perplexity 如何运作？，访问日期：2025年8月17日。 [在线]。获取地址：https://www.perplexity.ai/ help-center/en/articles/10352895-how-does-perplexity-work [42] D. B. Guruge、R. Kadel 和 S. J. Halder，“课程推荐系统方法学的最新进展——近期研究综述”，《数据》，第 6 卷，第 2 期，2021年。 [在线]。获取地址：https://www.mdpi.com/ 2306 5729 6 2 18 - / / / S. Shailendra、R. Kadel [43] 和 A. Sharma，“教育领域采用生成式人工智能（GenAI）的框架”，《*IEEE* 教育学报》，第 67 卷，第 5 期，页 777–785，2024年。[44] E. Kanoulas、V. Pavlu、K. Dai 和 J. A. Aslam，“对相关文档和不相关文档的分数分布建模”，载于信息检索理论会议。施普林格，2009年，页 152–163。[45] E. Kanoulas、K. Dai、V. Pavlu 和 J. A. Aslam，“分数分布模型：假设、直觉及对分数操纵的稳健性”，载于第 *33* 届国际 *ACM SIGIR* 信息检索研究与发展会议论文集，2010年，页 242–249。[46] S. Robertson，“论分数分布与相关性”，载于《信息检索进展》，G. Amati、C. Carpineto 和 G. Romano 编。柏林，海德堡：施普林格·柏林·海德堡，2007年，页 40–51。[47] A. M. Elmassry、N. Zaki、N. Alsheikh 和 M. Mediani，“自动作文评分中预训练模型的系统综述”，*IEEE Access*，第 13 卷，页 121 902–121 917，2025年。