使用大型语言模型进行临床综述的自动化论文筛选

Eddie Guo[1,*], Mehul Gupta[1], Jiawen Deng[2], Ye-Jean Park[2], Mike Paget[1], Christopher Naugler[1]
[1]加拿大卡尔加里大学卡明医学院，卡尔加里
[2]加拿大多伦多大学特梅蒂医学院，多伦多

*通讯作者：eddie.guo@ucalgary.ca，1-587-988-0292，加拿大阿尔伯塔省卡尔加里市 Hospital Dr NW 3330号，邮编 T2N 4N1

2023年5月2日

arXiv:2305.00844v1 [cs.CL] 1 May 2023

摘要

研究目的 评估 OpenAI GPT 应用程序接口在真实世界临床综述数据集中准确、高效地识别相关标题和摘要方面的表现，并将其表现与两名独立人工审阅人员的真实标注结果进行比较。

方法 我们提出了一种新颖的工作流程，使用 OpenAI GPT API 对临床综述中的标题和摘要进行筛选。我们创建了一个 Python 脚本，使用自然语言表述的筛选标准调用 GPT API，并输入经至少两名审阅人员筛选的标题和摘要数据集语料库。我们在六篇综述论文中，将模型的表现与人工审阅的论文进行比较，共筛选了超过 24,000 篇标题和摘要。

结果 我们的结果显示，准确率为 0.91，排除的论文的敏感度为 0.91，纳入的论文的敏感度为 0.76。在随机选取的部分论文中，GPT API 展现了为其决策提供推理依据的能力；当被要求解释其推理过程时，它还纠正了部分错误分类的初始判断。

结论 GPT API 有望简化临床综述流程，为研究人员节省宝贵的时间和精力，并提升临床综述的整体质量。通过优化工作流程并作为研究人员和审阅人员的辅助工具，而非取而代之，GPTAPI 能够提高效率，并帮助医学研究得出更准确、可靠的结论。

关键词： 摘要筛选、自然语言处理、系统综述、GPT

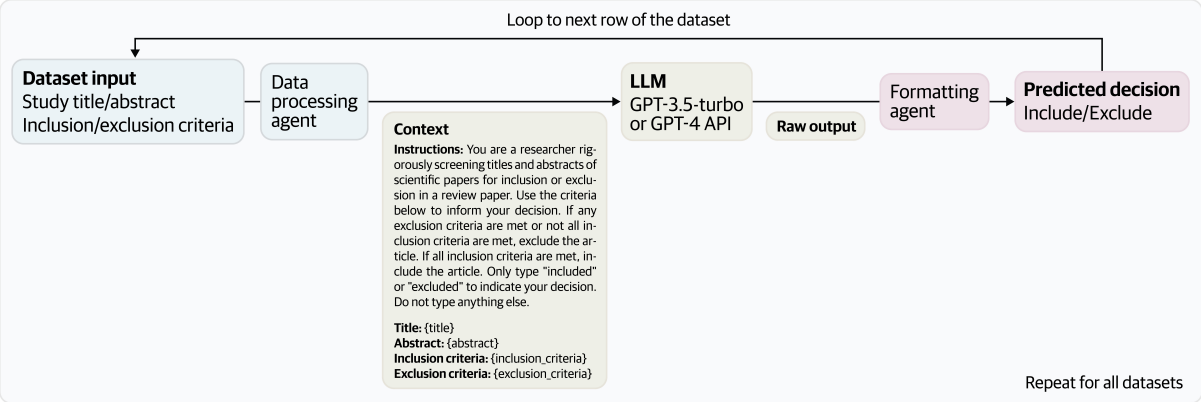
1 背景与意义

知识综合是整合并总结文献中相关研究、从而加深对某个主题理解的过程，是识别知识空白并为感兴趣主题的未来研究工作提供依据的关键环节 [1, 2]。系统综述和范围综述是多个学科中最常用且最严谨的知识综合形式之一 [1, 2]。鉴于系统综述和范围综述的结果能够为指南、方案和决策过程提供依据，尤其是对医疗保健领域的利益相关方而言，此类综述所呈现证据的质量可能产生重大影响 [3]。

系统综述和范围综述的质量在很大程度上取决于数据库检索的全面性以及后续的文献筛选流程。在这些关键步骤中遗漏相关文章可能会导致偏倚 [4], ，而纳入不一致的研究则可能得出误导性结论，并增加异质性[5]。因此，Cochrane手册等有关开展临床综述的指南 [6], 建议，文献筛选应由至少两名独立审阅人员分别完成。

然而，与单人筛选相比，双人筛选实际上会使完成系统综述所需的财力和人力资源增加一倍。对于规模较小的研究团队、纳入标准宽泛的综述项目（如网状荟萃分析），或时间受限的综述项目（如大流行初期有关COVID-19的综述）而言，这尤其成问题 [7]。此外，筛选决策往往存在显著的评审者间变异，因此需要花费额外时间讨论并解决分歧 [8]。由于时间紧迫以及资源浪费常伴随双人筛选，研究也可能采用更有针对性且更敏感的检索策略，导致检索过程中遗漏若干文章 [9]。此外，尽管每项研究的具体情况各不相同，许多系统综述可能会检索到数千篇文章，但在标题和摘要筛选阶段排除其中大多数（即多达 ~90%） [9, 10]。

人工智能和机器学习（机器学习）的最新进展，使系统综述工作流程中的重复性步骤得以半自动化或全自动化 [11, 12, 13]。此类应用的典型例子包括 RobotReviewer 和 TrialStreamer [14, 15], ，它们是机器学习模型，旨在从科学文章或摘要中提取信息，以评估研究质量并推断治疗效果。具体而言，研究表明，RobotReviewer（2016）评估偏倚风险的能力与人工审稿人相近，准确率仅相差约 7% [16]。同样，TrialStreamer 是一个旨在从全文中提取关键信息的系统，例如推断临床论文中哪些干预措施效果最佳，并比较某个主题下所有相关全文中提取的研究结局



Fig图 1：使用 GPT AP 自动筛选的 Python 脚本概览 I.

已编入 MEDLINE [17] 索引。

此前已有研究尝试将标题和摘要筛选流程自动化，但这些方法往往需要耗费大量人力或计算资源进行标注、预训练或向量化 [18]。例如，Rayyan 和 Abstracker 是两款免费的网络工具，通过自然语言处理算法学习审稿人在何时、何处纳入或排除文章，并据此模拟类似的筛选方式，从而半自动化地筛选文章 [19, 20]。Rayyan 还表现出较高的特异度：在该工具筛选完一项研究中待分析文章的 75% 后，所有相关文章中有 98% 已被纳入 [21]。虽然使用这些工具实现自动化被发现可以节省时间，但如果完全依赖工具自动筛选，仍存在从较低到实质性的漏掉研究的风险 [19, 20]。此外，目前的程序可能采用 n 元语法等传统标准方法，而非更先进的技术，例如生成式预训练变换器（GPT）模型。该模型使用通用领域的数据进行训练，无需额外训练即可学习嵌入，以表示词语相对于其他词语的语义和语境 [22, 23]。

本文中，我们介绍了一种新颖的工作流程：通过向公开可用的 OpenAI GPT API 提供通俗易懂的提示，筛选临床综述的标题和摘要。我们的目标是评估 GPT 模型能否从真实世界临床综述数据集中准确、高效地识别相关标题和摘要，以及能否解释其决策并反思错误分类。我们将模型的表现与两名独立人工审稿人对六篇综述论文所做的真实标注进行比较，共筛选了 24,000 多个标题和摘要。

2 方法

在本研究中，我们获取了一个由标题和摘要组成的语料库数据集，这些数据集已由至少两名人工审阅人员筛选，并用于训练我们的模型。随后，我们

编写了一个 Python 脚本，将每篇论文的筛选标准提供给 OpenAI GPT API。随后，我们使用统一的指令提示，将每篇论文提交给该 API，根据其标题和摘要内容判断是否应纳入或排除。我们以人工审阅的论文为参照，计算了总体准确率、纳入论文和排除论文的敏感度，以及通过科恩κ系数衡量的评审者间信度。所有数据和代码均可在 Mendeley Datasets 获取，数字对象标识符为：10.17632/np79tmhkh5.1。

2.1 数据收集

为验证我们提出的纳入/排除方法，我们从不同的系统综述和范围综述项目中获取了总计六个标题/摘要筛选数据集。这些项目涵盖了各种医学科学主题，在规模、方法和筛选标准的复杂程度方面各不相同（表1和补充表1）。我们获取了专家审阅者对每条标题/摘要记录作出的纳入/排除决定，以及筛选过程中提供给专家审阅者的标准。综述特征汇总见表2。

2.2 应用程序创建

给定一个数据集 `df_info`，其中包含关于纳入和排除标准的信息，以及待审查数据集中的标题和摘要，该应用程序调用 OpenAI GPT API，将每篇待筛选的论文分类为纳入或排除。该应用程序使用 Python 编写。提供给 GPT API 的提示词如下：

> **说明：** 你是一名研究人员，正在严格筛选科学论文的标题和摘要，以确定其是否纳入综述论文。请根据以下标准作出决定。如果符合任何一项排除标准，或未满足所有纳入标准，则排除该文章。如果满足所有纳入标准，则纳入该文章。只能输入“纳入”或“排除”来表示你的决定。不要输入任何其他内容。
>
> **摘要：** {摘要}
> **纳入标准：** {纳入标准}
> **排除标准：** {排除标准}
> **决定：**

其中，“**判定：**”表示 GPT API 是否纳入或排除该文章。因此，算法如下：

```
data_df <- load(df_info)
for each dataset in data_df:
    for each row in dataset:
```

表1：纳入的研究及其特征。前五个数据集为包含荟萃分析的系统综述。最后一项研究为范围综述。

研究标题 [数字对象标识符]	数据集 Name	n（纳入的 n）	研究类型	研究主题
疗效 and 安全性 的 IVER-伊维菌素 for the 治疗 新冠肺炎：一项系统综述 and 荟萃分析 [10.1093/qjmed/hcab247]	IVM	279（35）	系统综述 and 荟萃分析 随机和非随机试验	新冠肺炎治疗，抗疟药
疗效 and 安全性 of se-选择性 血清素 再摄取抑制剂 in 新冠肺炎管理：a 系统综述 and 荟萃分析 [10.1016/j.cmi.2023年.01.010]	SSRI	3,989 (29)	系统综述 and 荟萃分析 随机和非随机试验	新冠肺炎治疗，抗抑郁药
疗效 洛匹那韦–利托那韦 联合 治疗 for the 治疗 of 住院新冠肺炎患者：一项荟萃分析 [10.2217/fvl-2021-0066]	LPVR	1,456 (91)	系统综述 and 荟萃分析 随机试验和非随机试验	新冠肺炎治疗，抗逆转录病毒药物
针灸的使用 患者 with 雷诺现象综合征：A 系统综述-view and 荟萃分析 of 随机 对照 试验 [10.1177/09645284221076504]	雷诺氏	942 (6)	系统综述 and 荟萃分析 随机对照试验	雷诺现象综合征，针灸
比较 疗效 of ad-辅助 非阿片类 镇痛 成人心脏手术中 pa-患者：一项网状荟萃分析 [10.1053/j.jvca.2023.03.018]	NOA	14,771 (354)	系统综述以及网络荟萃分析-随机效应分析-化对照试-als	术后疼痛，镇痛药
评估研究现状 以及大语言模型在临床.....中的实用性和效用 临床环境：范围综述方案-范围综述*	LLM	2,870 (23)	范围综述	机器学习-ing in 临床医学
总计	-	24,307 （538）	-	-

*已在开放科学框架（OSF）注册，尚未发表：https://osf.io/498k6。

```
    prompt <- instructions + title + abstract + inclusion criteria \
        + exclusion criteria
    decision <- GPT(prompt)
    row['decision'] <- decision
save(dataset)
```

表 2：使用 GPT API 自动筛选的 Python 脚本的数据格式。分析前已移除所有非英语字符。

Data	列
df_info	数据集名称（str）：数据集的名称 纳入标准（str）：筛选纳入标准 排除标准（str）：筛选排除标准
数据集 *	标题（字符串）：论文标题 摘要（字符串）：论文摘要

*数据集的名称必须与 df info 中的 Dataset Name 一致。_

2.3 评估与数据分析

在该应用运行于我们分析中包含的所有数据集后，计算了以下指标：绝对一致性、决策标签的敏感度，以及 scikit-learn 返回的分类报告。

选取部分结果，让 GPT 解释其推理过程。将以下提示附加到提供给应用程序接口的原始提示开头：“请根据以下信息解释你作出该决定的推理过程。”随后将人工和 GPT 的决定附加到提示末尾。选取部分错误结果，让 GPT 反思其错误答案。将以下提示附加到提供给应用程序接口的原始提示开头：“请根据以下信息解释为何所作决定有误。”随后将人工和 GPT 的决定附加到提示末尾。结果列于表4中。

3 结果

GPT的总体准确率为0.91，纳入的论文的敏感度为0.76，排除的论文的敏感度为0.91（表 [结果], 图 [混淆矩阵]）。在 NOA数据集（n=14，771篇摘要，其中354篇纳入）上，模型运行了643m 50.8s，成本约为25 美元。数据集特征详见表1，模型性能见表3并在图2中可视化，GPT的推理汇总于表4。

表3：GPT 根据人工审稿人的判定结果筛选标题和摘要的表现。卡帕系数（人工）表示两位独立人工审稿人之间的一致性。卡帕系数（筛选）表示 GPT 与各数据集中最终纳入或排除的论文之间的一致性。

数据集	准确率	敏感度（纳入）	敏感度（已排除）	卡帕系数（人工）	卡帕系数（筛查）
IVM	0.748	0.686	0.756	0.72	0.26
SSRI	0.846	0.966	0.949	0.58	0.21
LPVR	0.949	0.593	0.862	0.51	0.25
雷诺氏	0.965	0.833	0.966	0.91	0.22
NOA	0.895	0.782	0.898	0.35	0.23
LLM	0.943	1.000	0.942	0.69	0.21
总计（加权平均）	0.907	0.764	0.910	-	-

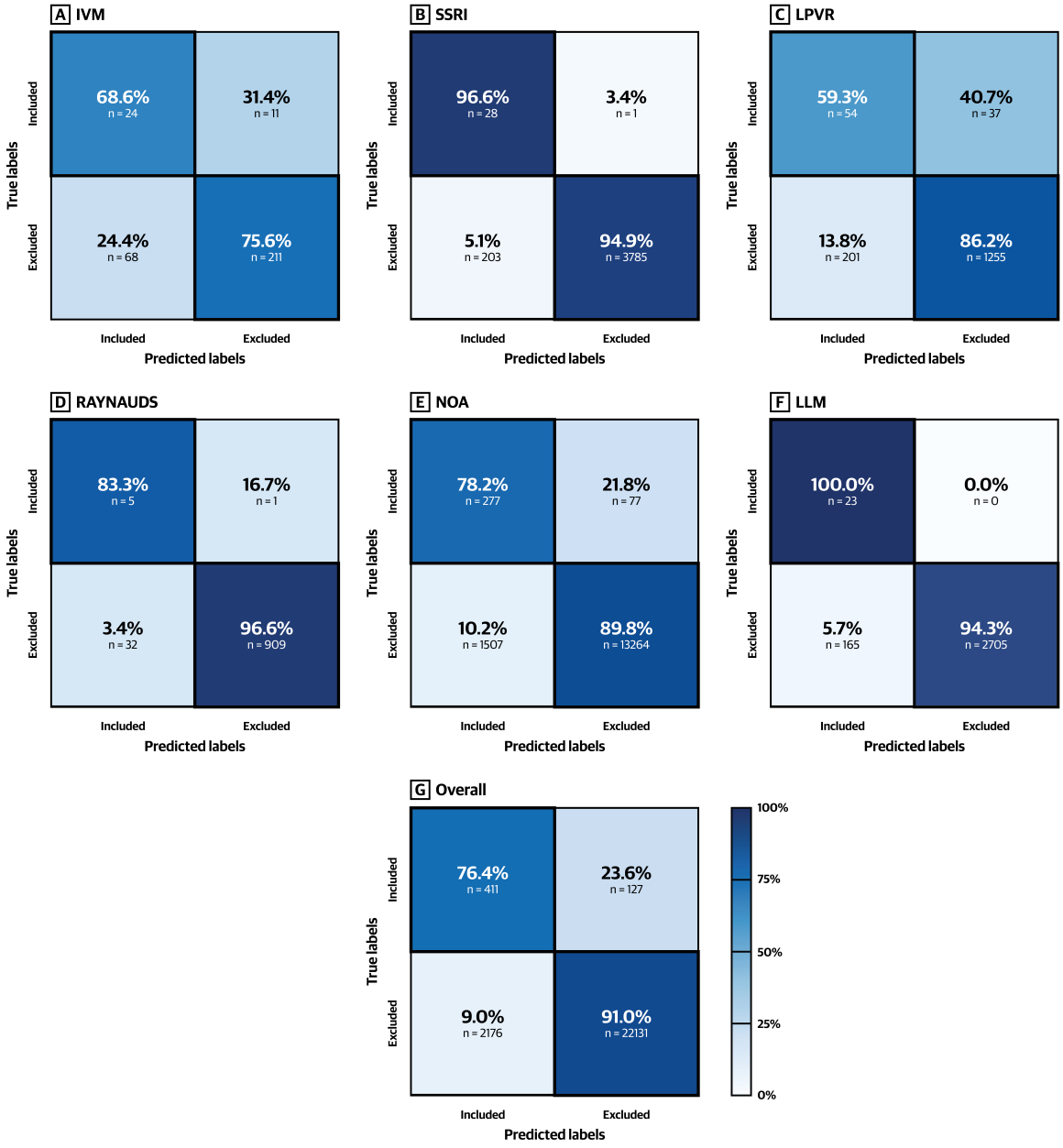


图 2：所有数据集（A-F）中纳入和排除论文的混淆矩阵，以及整体性能（G）。

Table 4: Decision and reasoning of the GPT API for a subset of four example titles and abstracts. Full abstracts are in IVM (Table 1).

Prompt	Decision and Reasoning
Title: Use of ivermectin in the treatment of Covid-19: A pilot trial **Abstract:** Objectives: In this randomized open-label trial pilot study we assessed the antiviral effects and safety of various doses of ivermectin in patients with mild clinical symptoms of COVID-19. Method(s): Patients were randomly assigned to receive standard of care (SOC) treatment at hospital admission...	**Human:** Included **GPT:** Included **Reasoning:** The study meets all of the inclusion criteria. It is a randomized controlled trial that compares ivermectin against standard of care treatment in adult COVID-19 patients. The study reports outcomes related to time to viral clearance and adverse events. Additionally, the study is published in a peer-reviewed journal. Therefore, the article should be included in the review paper.
Title: What does the COVID-19 pandemic mean for the next decade of onchocerciasis control and elimination? **Abstract:** Background: Mass drug administration (MDA) of ivermectin for onchocerciasis has been disrupted by the coronavirus disease 2019 (COVID-19) pandemic. Mathematical modelling can help predict how missed/delayed MDA will affect short-term epidemiological trends and elimination prospects by 2030...	**Human:** Excluded **GPT:** Excluded **Reasoning:** The article does not meet any of the inclusion criteria or exclusion criteria listed. It is not a randomized controlled trial or comparative cohort study, it does not compare ivermectin against standard of care or control group/placebo, it does not include adult COVID-19 inpatients and/or outpatients, and it does not report any of the specified outcomes. Additionally, there is no indication that the article is non-peer-reviewed. Therefore, the article is excluded.
Title: Use of Ivermectin Is Associated With Lower Mortality in Hospitalized Patients With Coronavirus Disease 2019: The Ivermectin in COVID Nineteen Study **Abstract:** Background: Ivermectin was shown to inhibit severe acute respiratory syndrome coronavirus 2 replication in vitro, which has led to off-label use, but clinical efficacy has not been described previously. Research Question: Does ivermectin benefit hospitalized coronavirus disease 2019 (COVID-19) patients?...	**Human:** Included **GPT:** Excluded **Reasoning:** The article meets all of the inclusion criteria and none of the exclusion criteria, therefore it should be included in the review paper. **Reflection:** The given decision of excluding the article is incorrect. The article meets all the inclusion criteria as it is a non-randomized comparative cohort study that compares ivermectin against standard of care or control group/placebo, includes adult COVID-19 inpatients, and reports mortality incidence. The article does not use ivermectin for prophylaxis of COVID-19 and is a peer-reviewed article. Therefore, the correct decision should be "Included".
Title: What to expect from different drugs used in the treatment of COVID-19: A study on applications and in vivo and in vitro results **Abstract:** The end of 2019 was marked by the emergence of a new type of coronavirus (SARS-CoV-2), which has killed more than 240,000 people...	**Human:** Excluded **GPT:** Included **Reasoning:** The study is not a randomized controlled trial or a non-randomized comparative cohort study, which is an inclusion criteria. Therefore, the article does not meet all inclusion criteria and must be excluded. **Reflection:** As per the given inclusion and exclusion criteria, the decision given is incorrect. The abstract does not mention the use of ivermectin in any of the studies. Therefore, the article cannot be included based on the inclusion criteria.

4 讨论

在本研究中，我们评估了 OpenAI GPT API 在临床综述论文纳入/排除标准选择中的表现。数据显示，该应用的总体准确率为 0.91，表明其决策与参考标准之间具有较高的一致性。对于纳入论文，应用的敏感度为 0.76，说明其正确识别相关论文的表现中等（表3，图2）。对于排除论文，敏感度为 0.91，表明该应用能够有效排除不相关论文。这些结果凸显了 GPT API 支持临床综述流程的潜力。

4.1 GPT API 在综述流程中的表现所带来的启示

GPT API 的表现对临床综述论文纳入/排除标准选择的效率和一致性具有多方面的影响。通过优化工作流程，并作为研究人员和审阅人员的辅助工具而非替代者，GPT API 有望简化综述流程。效率的提升能够为研究人员和临床医生节省宝贵的时间和精力，使他们得以专注于更复杂的任务和深入分析。此外，该应用程序接口无需用户预先训练，并且能够说明其纳入或排除论文决策的理由；而传统自然语言处理算法在自动化或半自动化论文筛选中缺乏这一能力（表4）。值得注意的是，当被要求解释部分错误分类的理由时，GPT API 修正了其最初的判断。最终，效率提升与推理能力相结合，有望提高临床综述的整体质量，从而使医学研究的结论更加准确、可靠。

在综述流程中使用 GPT API 还有助于提高相关论文筛选的一致性。通过自动化流程中的某些环节，并为研究人员和临床医生提供辅助，该应用程序接口能够简化综述流程，减少人为错误和偏倚的可能性，从而得出更客观、可靠的结果 [24]。这种一致性的提升反过来还可能改善临床综述中综合证据的整体质量，为医学决策和临床指南的制定提供更坚实的依据。

在资源有限时，GPT API 作为决策工具的潜在价值尤为突出。在这种情况下，可将该 API 用作初步决策辅助工具，简化综述流程，并使人工筛选人员能够专注于数量较少且相关性更高的论文。通过自动化初始筛选流程，该 API 有助于减轻研究人员和临床医生的工作负担，使他们能够更高效地分配

时间和精力。

在初筛阶段将 GPT API 用作第一轮决策辅助工具，还有助于降低人为错误和偏倚的风险，从而促进更加客观、一致的论文筛选。尽管该 API 纳入相关论文的敏感度可能并不完美，但其在排除不相关论文方面的高特异度，仍可为缩小潜在相关研究的范围提供有价值的支持 [9]。当需要筛选大量论文而人力资源又十分有限时，这一点尤其有益 [25]。

4.2 在综述流程中实施 GPT API 的局限性与挑战

尽管 GPT API 在简化综述流程方面展现出良好前景，但有必要承认其局限性与挑战。一个显著的局限是，其排除论文时的高特异度为 0.91，而纳入论文时的敏感度较低，仅为 0.76。这一差异表明，尽管该 API 能够有效排除不相关论文，但在识别应纳入的相关论文方面可能并不那么擅长。这可能导致综述流程遗漏重要研究，进而影响最终综述的全面性和质量。

因此，不应将 GPT API 视为人工专业知识的替代品，而应将其视为一种辅助工具，以提升综述流程的效率和一致性。人工筛选人员仍应参与最终决策，尤其是在该应用程序接口纳入相关论文的敏感度可能不足的情况下 [7]。将 GPT API 的优势与人工专业知识相结合，研究人员便能优化综述流程，并确保最终综述的准确率和全面性。

4.3 未来研究与发展

未来研究与开发可从多个方向展开，包括优化 GPT API 在临床综述论文场景中的表现，纳入研究类型和年份等元数据，并探索少样本学习方法。此外，通过微调训练生成器—判别器模型，或可提升该 API 的表现 [26]。还可以探索将 GPT API 扩展应用于医学研究或文献综述的其他领域。这将涉及使用大语言模型（LLM）执行诸如识别和提取研究设计信息、患者特征及不良事件等任务。随着未来版本的 GPT 模型最大令牌长度增加，筛选整篇论文或将成为可能 [27]。此外，探索使用大语言模型生成临床综述论文，也可能是一个有前景的研究方向。

5 结论

GPT API 显示出潜力，可作为提高临床综述论文纳入/排除标准筛选效率和一致性的有价值工具。尽管其实施仍存在局限与挑战，但本研究中的表现表明，它可能对临床综述论文撰写和医学研究产生更广泛的影响。未来的研究与开发应着重优化该 API 的表现、拓展其应用范围，并探索其在临床研究其他方面的潜力。

致谢

我们谨向以下专家审阅者表示感谢，感谢他们为本研究所使用的综述数据集中提供筛选判定，并同意将这些数据集公开发布：Abhinav Pillai、Mike Paget、Christopher Naugler、Kiyan Heybati、Fangwen Zhou、Myron Moskalyk、Saif Ali、Chi Yi Wong、Wenteng Hou、Umaima Abbas、Qi Kang Zuo、EmmaHuang、DanielRayner、CristianGarcia、HarikrishnaaBa Ramaraju、Oswin Chang、Zachary Silver、Thanansayan Dhivagaran、Elena Zheng、Shayan Heybati。

利益冲突

作者声明不存在利益冲突。

资金支持

本研究未获得资助。

CRediT 作者贡献声明

Eddie Guo：概念构思、方法学、软件、正式分析、调查、撰写——初稿、撰写——审阅与编辑、可视化、监督、项目管理；**Mehul Gupta：**概念构思、方法学、调查、撰写——初稿、撰写——审阅与编辑、监督、项目管理；**Jiawen Deng：**方法学、软件、正式分析、调查、数据整理、撰写——初稿、可视化；**Ye-Jean Park：**方法学、正式分析、调查、数据整理、撰写——初稿、可视化；**Mike Paget：**撰写——审阅与编辑；**Christopher Naugler：**撰写——审阅与编辑

参考文献

[1] Sargeant JM, O'Connor AM. 范围综述、系统综述与荟萃分析：在兽医学中的应用。*Front Vet Sci* 2020年;7:11.[2] Garritty C, Stevens A, Hamel C, 等. 循证医学中的知识综合。*Semin NuclMed* 2019年;49(2):136–144.[3] Luchini C, Veronese N, Nottegar A, 等. 评估荟萃研究中研究的质量：关于最重要质量评估工具的综述/指南。*PharmStat* 2021年;20(1):185–195.[4] Gartlehner G, Affengruber L, Titscher V, 等. 单人审阅者进行摘要筛选时漏掉了13%的相关研究：一项基于众包的随机对照试验。*JClinEpidemiol* 2020年;121:20–28.[5] Fletcher J. 什么是异质性？它重要吗？*BMJ* 2007年;334(7584):94–96.[6] Higgins JPT, Thomas J, Chandler J, 等，编者。干预措施系统综述Cochrane手册（v6.2）。Cochrane协作网，2021年。[7] Chai KEK, Lines RLJ, Gucciardi DF, 等. Research Screener：一种用于半自动化系统综述摘要筛选的机器学习工具。*Syst Rev* 2021年; 10(1):93.[8] Tuijn S, Janssens F, Robben P, 等. 减少评审者间变异并改善医疗保健：荟萃分析综述。*J Eval Clin Pract* 2012年; 18(4):887–895.[9] Rathbone J, Carter M, Hoffmann T, 等. 为系统综述审阅人员提供更好的重复项检测：Systematic Review Assistant-Deduplication Module评估。*SystRev* 2015年;4(1):6.[10] Polanin JR, Pigott TD, Espelage DL, 等. 大型证据系统综述和荟萃分析摘要筛选的最佳实践指南。*Res Synth Methods*2019年;10(3):330–342.[11] Marshall IJ, Wallace BC. 迈向系统综述自动化：在研究综合中使用机器学习工具的实用指南。*Syst Rev* 2019年;8(1):163.[12] Marshall IJ, Trikalinos TA, Soboczenski F, 等. 一项试点研究表明，自动化实时系统综述更新可行、准确且节省工作量。*J ClinEpidemiol* 2023年;153:26–33.[13] Blaizot A, Veettil SK, Saidoung P, 等. 在健康科学系统综述中使用人工智能方法：一项系统综述。*Res Synth Methods* 2022年; 13(3):353–362.

[14] Marshall IJ, Kuiper J, Banner E, 等. 自动化生物医学证据综合：RobotReviewer. 计算语言学协会会议论文集2017;2017:7–12.[15] Marshall IJ, Nye B, Kuiper J, 等. Trialstreamer：一个持续更新的临床试验报告数据库。美国医学信息学会杂志2020年;27(12):1903–1912.[16] Marshall IJ, Kuiper J, Wallace BC. RobotReviewer：评估一种自动评估临床试验偏倚的系统。美国医学信息学会杂志2016年;23(1):193–201.[17] Nye B, Nenkova A, Marshall I, 等. Trialstreamer：实时映射和浏览医学证据。载于：计算语言学协会第58届年会论文集：系统演示。计算语言学协会，线上，63–69.[18] Moreno-Garcia CF, Jayne C, Elyan E, 等. 机器学习和零样本分类方法在系统综述自动摘要筛选中的一种新颖应用。决策分析期刊2023年;6(100162):100162.[19] Wallace BC, Trikalinos TA, Lau J, 等. 面向系统综述的生物医学引文半自动筛选。*BMC*生物信息学2010年;11:55.[20] Ouzzani M, Hammady H, Fedorowicz Z, 等. Rayyan——一款用于系统综述的网页和移动应用。*Syst Rev* 2016年;5(1):210.[21] Olofsson H, Brolund A, Hellberg C, 等. 能否通过文本挖掘减少摘要筛选工作量？Rayyan工具的用户体验。*Res Synth Methods*2017年;8(3):275–280.[22]Shree P. GPT模型解析。Open AI的GPT-1、GPT-2、GPT-3。`https://medium.com/walmartglobaltech/the-journey-of-open-ai-gpt-models-32d95b7b7fb2`，2020年。访问日期：2023年4月19日。[23] O'Mara-Eves A, Thomas J, McNaught J, 等. 在系统综述中使用文本挖掘识别研究：当前方法的系统综述。*Syst Rev*2015年;4(1):5.[24] Zhang Y, Liang S, Feng Y, 等. 使用机器学习自动筛选文献以综合医学证据：诊断性试验准确率系统综述方案。*SystRev* 2022年;11(1):11.[25] van de Schoot R, de Bruin J, Schram R, 等. 一种用于高效、透明系统综述的开源机器学习框架。自然机器智能2021年;3(2):125–133。

[26] 微调分类器以提高真实性。https://help.openai.com/en/articles/5528730-fine-tuning-a-classifier-to-improve-truthfulness。访问日期：2023年4月20日。

[27] GPT-4 各模型之间有什么区别？ https://help.openai.com/en/articles/7127966-what-is-the-difference-between-the-gpt-4-models。访问日期：2023年4月19日。