ADVISE：人工智能辅助加速全球发展领域证据综合的设计

Kristen M. Edwards[1], †,*, Binyang Song[1]†, Jaron Porciello 2, Mark Engelbert3, Carolyn Huang3, Faez Ahmed[1]

[1]麻省理工学院，马萨诸塞州剑桥[2]圣母大学，印第安纳州
南本德[3]国际影响评估倡议公司

arXiv:2305.01145v1 [cs.CL] 2 May 2023

摘要

在制定循证政策和项目时，决策者必须从庞大且快速增长的文献库中提炼关键信息。识别相关文献。

f 从原始搜索结果中筛选文档既耗时又耗费资源，通常需要人工筛选。本研究开发了一种基于双向变换器编码器表示（*BERT*）模型的人工智能代理，并将其纳入一个人类团队，共同设计用于全球发展的证据综合产品。我们探究了人机混合团队加速证据综合过程的有效性。为进一步提高团队效率，我们通过主动学习（*AL*）增强人机混合团队。具体而言，我们探究了不同的抽样策略，包括随机抽样、最低置信度（*LC*）抽样和最高优先级（*HP*）抽样，以研究这些策略对协作筛选流程的影响。结果表明，在识别全部相关文档中*80%*的文档时，与没有人工智能辅助的情况相比，将基于*BERT*的人工智能代理纳入人类团队可减少*68.5%*的人工筛查工作量；与使用基于支持向量机（*SVM*）的人工智能代理的情况相比，可减少*16.8%*。当我们采用最高优先级抽样策略时

f 采用主动学习时，与不使用人工智能辅助相比，在识别出全部相关文档的 *80%* 时，人工筛查工作量还可进一步降低至 *78.3%*。我们将主动学习增强的人机混合协作工作流应用于美国国际开发署三个证据缺口图的设计流程中，并发现其效果显著。这些

fi 研究结果展示了在人机混合协作环境下，人工智能如何加速证据综合产品的开发，并促进全球发展领域及时开展循证决策。

关键词：设计中的人工智能、自然语言处理、全球发展、证据综合

1. 引言

2011年，美国国际开发署（美国国际开发署）发布了《评估政策》，并由此作出一项雄心勃勃的承诺：严格评估证据，以制定循证政策 [1]。循证政策是指以经过评估的客观证据为基础或依据的公共政策。为强调循证政策在美国国际开发署和美国政府中的重要性，《2018年循证政策制定基础法》要求该法案所涵盖的所有机构“确认本机构致力于开展严格且相关的评估，并利用评估证据为政策和实践提供依据” [2]。这些政策之所以至关重要，部分原因在于它们决定着数十亿美元的支出。例如，2017年，美国国际开发署仅用于对外农业援助就支出了10.1亿美元 [3]。

然而，当前的信息爆炸使评估所有可用证据变得十分繁重。仅在2018年，科学与工程领域的全球研究产出就达到260万篇文章；从2008年到2018 [4]，其年增长率为4%。个人理解所有可用研究的能力是有限的。因此，政策制定者转而采用证据综合，以理解日益增长的研究成果，并作出明智的决策。证据综合是指汇总来自众多来源和学科的信息与知识，为决策提供依据的过程 [5, 6]。然而，创建证据综合产品（如证据缺口图（EGM））需要人类专家投入大量时间和精力。如“相关工作”部分所述，EGM将干预措施及其相关结果可视化 [7], ，并已被证明能够为从农业到公共卫生等领域的决策者带来巨大的价值 [5]。例如，图 1 展示了一个来自3ie的EGM局部内容 [1]。我们可以看到，在“水资源获取与管理”和“改良种子”等干预措施与“利润”等结果之间存在研究空白。政策制定者可以据此规划未来的投资和研究。

我们的目标是加快全球发展领域证据图谱的设计，并减轻信息筛选的负担。国际影响评估倡议（3ie）是全球生成用于决策的证据图谱的领先机构之一。3ie 当前的证据综合流程包括大量由专家对文档进行筛选，并适度运用机器智能，通常需要近六个月才能完成 [6]。自然语言处理（NLP）是人工智能的一种形式

†共同第一作者
*通讯作者

[1]https://developmentevidence.3ieimpact.org/egm/食品系统与营养-证据缺口图谱

（人工智能），长期以来一直用于文本理解。近年来，基于规则的自然语言处理模型受到了一些关注，并被用于促进医疗、法律和全球发展领域的循证决策 [8–10]。将最新的基于 Transformer 和迁移学习的自然语言处理模型融入其中，或可进一步改进已成功开展的相关工作。

1.1 贡献

标题和摘要（TA）筛选是证据差距图设计流程中最耗时的步骤之一，通常需要理解数万或数十万篇论文的标题和摘要并进行筛选。通过与 3ie 合作，我们做出了以下贡献：

1. 我们开发了一种基于BERT的人工智能代理，以加快证据差距图设计流程中的标题和摘要筛选环节，并将其纳入人工团队，以探究人机协作带来的效率提升。在最佳组合下，与没有人工智能辅助的情况相比，我们的人工智能代理在识别出所有符合条件文档的 80% 时，可将人工工作量减少 78.3%。

2. 我们将基于BERT的人工智能代理与行业标准的基于支持向量机的模型进行比较，发现基于BERT的模型在模型性能（三个证据差距图的平均准确率提高 12%）和节省工作量方面（模拟案例中所需工作量减少 17%，三个已部署证据差距图的工作量平均减少 46%）均优于支持向量机。

3. 我们确定了兼顾模型性能和节省人工工作量的最佳训练规模（5,000 份文档）。

4. 我们比较了主动学习策略，发现使用最高优先级或最低置信度策略，在识别出全部纳入文档的 80% 时，相较于不使用主动学习的 BERT，可额外减少 30% 的人工工作量。

5. 我们支持构建三个证据图谱：农业、营养和韧性。

2. 相关工作

以下各节将介绍证据缺口图、自然语言处理和主动学习领域的相关工作，尤其是人机团队方面的研究。

2.1 证据缺口图

EGM是证据综合的一种形式，即汇集来自多个来源和学科的信息与知识，为决策提供依据的过程 [5, 6]。通过系统地收集、分类和分析大量研究，证据综合能够比单项研究提供更可靠的主题信息 [11]。医学领域广泛推广了用于决策的证据综合，但它能为任何领域的决策者带来明显益处 [9, 12, 13]。因此，对于致力于制定政策和资助研究的全球发展领域决策者而言，证据综合是一项极具价值的工具 [5]。

国际影响评估倡议（3ie）率先采用了 EGM，即以可视化方式概览特定领域已完成和正在进行的影响评估与系统综述 [14]。3ie 通过以下方式创建这些 EGM：“[收集] 有关影响评估和系统综述的信息，以衡量国际发展政策和项目的影响” [7]。最终成果是一个矩阵，纵轴按“干预”类别划分，横轴按“结果”类别划分。干预是研究中采取的行动，结果则是该行动带来的影响。矩阵中的每个单元格都包含严格评估某项特定干预对某项特定结果影响的研究。

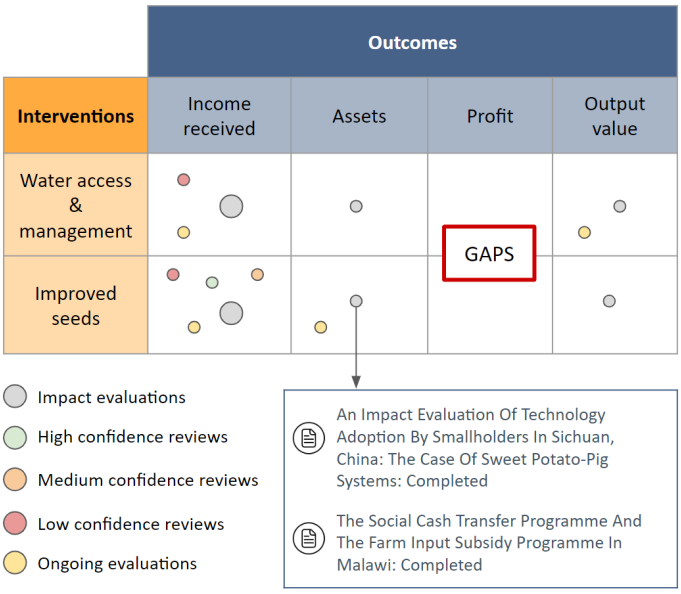


图1：3ie证据与差距图部分内容示意图，展示了两项干预措施和四种结果。两项干预措施与“利润”这一结果之间存在研究空白。不同颜色的点代表不同的证据类型。点的大小表示各组中文档的数量。

3ie 为 EGM 制定了全球标准，绘图方法也已被包括坎贝尔协作组织、世界银行独立评估局和美国国际开发署在内的机构采用 [14]。与其他形式的证据综合一样，EGM 首先会全面、系统地检索学术数据库和“灰色文献”来源（例如政府文件存储库或智库网站），以找出可能相关的研究。随后，证据与差距图团队会筛查检索结果，找出符合 EGM 标准的研究，这些标准涉及所评估的干预措施、所衡量的结果、实施环境和研究设计。确定符合条件的研究后，证据与差距图团队会提取每项研究中有关干预措施、结果及其他关键特征的信息，以确定其在证据与差距图矩阵中的位置，并分析文献趋势。

3ie 使用一款名为 EPPI-Reviewer 的软件来协助创建 EGM。虽然 EPPI-Reviewer 具备一些机器学习功能，可以加快筛选速度，但 [15], 大多数 EGM 任务仍由人工完成。因此，每个 EGM 都需要投入大量人工工作量和专业知识，许多 EGM 几乎需要六个月才能完成 [6]。鉴于政策制定者使用证据的主要障碍之一是缺乏及时的

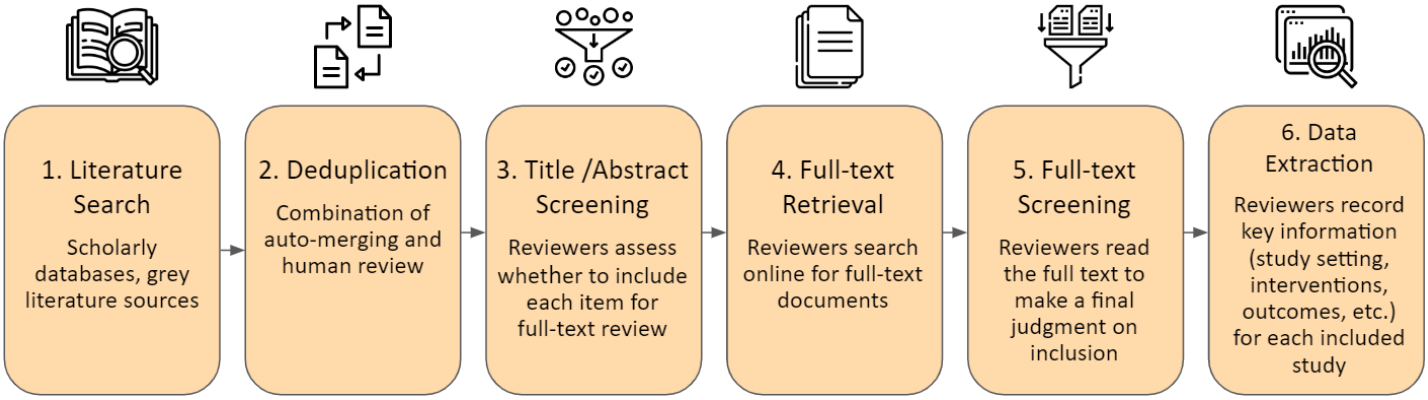


图2：当前证据图谱创建流程概览。

研究成果 [16], 因此，迫切需要减少完成 EGM 设计和开发过程所需的时间与工作量。

图中展示了设计证据图谱的主要步骤2。我们的工作聚焦于第三步：审阅者根据文档的标题和摘要筛选文档，以决定是否将其纳入证据图谱。入选文档将进入全文审查阶段。我们构建了三个基于 Transformer 的自然语言处理模型，可在此步骤自动对文档进行纳入分类。

2.2 证据综合中的自然语言处理

自然语言处理是机器学习的一个领域，在该领域中，计算机经过训练后能够理解文本和口语。传统上，在自然语言处理中，词语被表示为向量，相似的词语在连续空间中彼此相近 [17–19]。在医学领域，一种名为 BioMedICUS、用于自动化证据综合的自然语言处理模型提升了生物医学和临床报告文本分析与处理的可扩展性及性能 [8]。自然语言处理在医学领域的成功促使其应用扩展到其他领域，例如 LexNLP 等模型可自动从法律文本中提取信息 [9]。

目前有多种行业标准的自然语言处理工具用于协助人类专家设计证据综合产品，例如证据图谱。最常用的工具包括 EPPI-Reviewer、Rayyan 和 RobotReviewer [12, 13], , 所有这些工具都将支持向量机（SVM）作为其主要机器学习模型 [20–22]。

此外，基于规则的自然语言处理模型也已在多个领域得到探索，以促进循证决策 [10]。然而，基于规则的模型通常针对特定案例。将一个证据图谱中的基于规则的模型调整到另一个证据图谱，需要付出大量努力。此外，要全面捕捉人类使用的所有主观标准，并将其嵌入预先定义的规则中，也颇具挑战性。

现代自然语言处理模型的发展在很大程度上受到2017年 Transformer 引入的影响。Transformer 使文本输入能够并行处理，并在许多自然语言处理任务中取得了优于支持向量机及其他模型的先进成果 [23]。双向变换器编码器表示（BERT）是最知名的 Transformer 模型之一，并已在语言翻译和问答等自然语言处理任务中得到广泛研究任务 [24, 25]。其他此类模型还包括 GPT 及其衍生模型 [26]。

我们的工作探索了基于 BERT 的自然语言处理模型，将其作为人机团队设计面向全球发展的证据图谱的工具；与其他领域相比，全球发展涉及更多非结构化研究和更广泛的领域。与我们工作最为相近的近期研究可能是 srBERT [27], , 该研究使用特定主题的文章对 BERT 模型进行微调，以加快针对“艾灸改善认知障碍”的系统综述的筛选过程 [27]。另一方面，我们的工作使用规模更大、范围更广的数据集来创建证据图谱。此外，我们还在一个真实的人机团队中展示了自然语言处理工具的有效性，并最终创建了三个已部署的证据图谱，分别应用于农业、营养和韧性领域。我们利用创建已部署证据图谱的经验，探索设计流程中人机协作的细微之处。

2.3 主动学习与人机团队

在许多人工智能任务中，获取带标签的训练数据成本高昂且耗时 [28]。我们希望探索通过主动学习（主动学习, AL）减少所需训练数据规模的方法。主动学习的概念是：如果允许机器学习（ML）算法自行选择学习所用的数据，那么即使训练数据较少，也能取得更好的表现。主动学习已应用于图像分类 [29, 30], 、语音识别 [31], 、数据窃取检测 [32], 以及许多自然语言处理（NLP）任务 [33]等深度学习问题。学习器可能能够提出查询的主要问题设置或场景有三种：成员查询合成 [34], 、基于流的选择性采样 [35], 以及基于池的采样[36]。最常用的基于池的采样策略会根据信息量评估并排列整个未标注数据池，然后选出最佳查询 [37]。此外，还有不同的查询策略用于选择要查询的未标注实例。最常用的查询策略是不确定性采样 [36]。在这种策略中，学习器会查询那些它最不确定如何标注的实例 [37]。在不确定性采样类别中，有三种主要指标用于评估学习器对每个实例的不确定程度：最低置信度 [38], 、间隔采样 [39],和熵 [40]。

在训练过程中，主动学习试图优化信息

从人类流向人工智能，以便用更少的训练数据提升人工智能的性能。在本研究中，人工智能与人类协同工作，而非独立工作。因此，从人类流向人工智能以及从人工智能流向人类的信息流，对人机混合团队的表现都至关重要。因此，在这种情境下，主动学习应考虑双向信息流。

事实上，阻碍人工筛选人员将人工智能纳入其证据缺口图设计流程的常见障碍包括现有工作流程不匹配，以及学习曲线陡峭 [13]。研究表明，在真实任务中使用主动学习的人机团队，如果人工智能代理在训练过程中给出无关的建议或预测，可能会失去人类代理的信任 [32]。因此，我们探讨并讨论如何将人工智能工具有效整合到现有的证据缺口图设计流程中。我们还采用基于主动学习的方法，在最大限度提高人工智能分类器准确率的同时，尽可能减轻人工筛选人员的工作负担。

3. 方法论

在本研究中，我们使用基于BERT的自然语言处理模型来加快证据缺口图的设计流程。我们将这一自然语言处理工具应用于由3ie成员组成的真实人机团队的工作流程中，并支持了农业、营养和韧性领域三张已部署证据缺口图的开发。

我们分析不同机器学习方法和训练集规模如何影响人机团队的表现，重点关注模型准确率与人工工作量之间的权衡。我们将微调后的BERT模型与行业标准的、基于支持向量机的自然语言处理工具进行比较。此外，我们还探讨了采用不同查询策略的主动学习对模型准确率和人工工作量的影响。

我们的工作由两个案例研究组成：

1.**已部署的 EGM 设计**：主动设计并创建三个关于农业、营养和韧性的 EGM。

2. **模拟 EGM 设计**：利用现有的、已完整标注的数据集，回顾性研究用于创建 EGM 的最有效分类算法（支持向量机与 BERT）和主动学习策略。

这两个案例研究面临不同的挑战，优先事项也各不相同。在模拟 EGM 设计中，我们拥有完整标注的数据集，因此可以在其上尝试多种技术。而在已部署的 EGM 设计中，我们身处真实情境，需要借助人机团队从头开始创建 EGM。因此，我们只有那些经过专门选择进行筛选的文档才有标签。

此外，在已部署的 EGM 设计中，我们的目标是设计出尽可能全面、信息丰富的 EGM，同时高效利用人力资源。因此，我们希望尽量减少人类专家筛选无关文档所花费的时间，只筛选相关文档。这与传统主动学习的策略形成对比：传统策略会查询或筛选我们最不确定的文档。

3.1 数据集描述

对于模拟 EGM 设计，我们的数据由 3ie 提供，源自 3ie 手工标注文档中的发展-展证据门户（DEP）[2] [41], 这是一个涵盖全球发展领域众多领域的影响评估和系统综述的庞大存储库。我们使用在 3ie 的 DEP 中经过纳入筛选的 68,539 份文档数据集，来开发和评估我们的分类模型。表1 展示了该数据集的关键属性，例如标题、摘要和纳入决定。

属性	描述
标题	论文标题。
摘要	论文摘要。
关键词	论文关键词。
Year	出版年份。
出版类型	期刊、会议论文集、re-端口等。
来源	论文来源，例如，期刊-期刊或会议。
纳入决策	是否将该论文纳入为一项相关研究。如果不是，什么是排除标准。

表1：发展证据门户数据集中的关键属性。

在本研究中，每篇论文的标题都被整合到摘要中，作为开头的一句话。BERT分类模型以整合后的文本为输入。“纳入”或“排除”标签根据纳入决策确定。为了训练二元分类模型，“0”类对应“排除”的论文，“1”类则包含“纳入”的论文。该数据集高度不平衡，包含5,281篇纳入论文和63,258篇排除论文。排除论文的标准也被提取出来，用于训练特定标准分类模型。

对于已部署的 EGM 设计，我们正在积极设计三个 EGM。如图所示2，根据 3ie 的 EGM 工作流程，我们通过在学术数据库和灰色文献来源中进行文献检索来收集初始数据集。三个 EGM 的初始数据集规模分别为：农业 221k、营养 117k、韧性 60k。

3.2 数据预处理

对原始文档进行预处理以去除噪声。在此步骤中，会去除两类噪声。第一类是非英语文本。部分论文以多种语言提供标题和摘要。由于我们的模型只接受英语文本作为输入，因此其他语言的句子对模型而言属于噪声，应予以删除。第二类是与文档范围无关的英语文本内容，例如版权声明。预处理包括五个步骤。(1) 将每篇文档解析为句子。(2) 使用语言检测模型识别以非英语书写的句子。(3) 我们人工为 500 篇文档中的句子标注“相关”和“不相关”标签。(4) 使用标注数据训练 BERT分类模型，以预测其他句子的标签。该模型的准确率高于

[2]https://developmentevidence.3ieimpact.org/

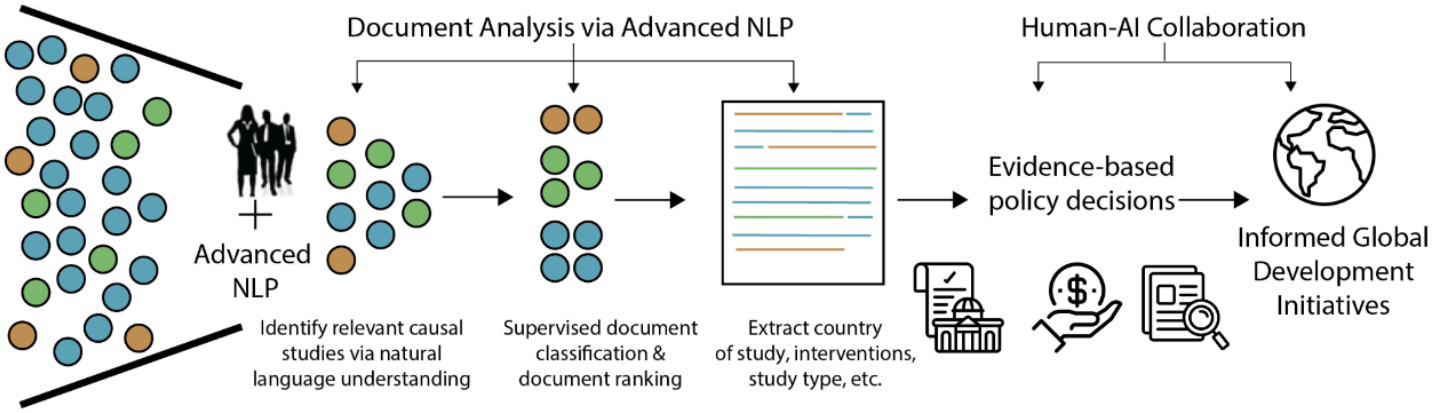


图 3：拟议在人机团队中运用自然语言处理工具，对文档（以圆圈表示）进行筛选、理解和分类，以便为循证政策决策提供依据。我们的目标是加速全球发展领域证据图谱产品的设计流程。

0.99。(5) 删除不相关的句子后，将剩余的相关句子整合回原始文档。

3.3 优先级评分

在本研究中，人工智能代理通过 BERT 二元分类模型实现，该模型采用一个具有 12 层、预训练且不区分大小写的 BERT 嵌入模块，其隐藏层大小为 768。BERT 嵌入模块之后连接一个丢弃率为 =0.1 的 dropout 层，以及一个输出二维（2D）向量作为最终分类预测结果的线性层。如上所述，人工智能代理需要根据分类模型预测的论文相关概率，对未标注论文进行采样或优先排序。该概率在定义 1 中被命名为“优先级分数”（PS）。

定义 1. 优先级评分 是人工智能代理预测论文为相关论文的概率，计算方式为$PS(p) = softmax(Pred(p))$ [1],其中，$Pred(p)$ 是分类模型对论文 p 的预测输出，它是一个二维向量。公式中的“1”表示 PS(p) 是论文被归类为“1”类的概率。根据这一定义，预测为相关论文的概率越高，其筛选优先级评分也越高。

3.4 采样策略

根据预测的优先级评分，我们采用三种不同的查询策略，从未筛选子集中抽取论文，并在下一轮中对其进行标注，然后将其加入训练集。

1. **最低置信度**： 最低置信度（*LC*）查询策略是主动学习中常用的策略之一，它会采样模型最不确定如何分类的论文 [37], ，如

$$x^*_{LC} = \arg\max_{x \in X} U(x) \tag{1}$$

论文的分类不确定性 $U(p)$ 由分类模型输出 $Pred(p)$ 通过以下方式得到

$$U(p) = 1 - max(softmax((Pred(p)))) \tag{2}$$

2. **最高优先级**：最高优先级（*HP*）查询策略抽取优先级评分最高的论文，公式如下：

$$x^*_{HP} = \arg\max_{x \in X} PS(x) \tag{3}$$

此查询策略改编自不确定性采样策略 [36]。对于证据综合，所有相关论文都需要由人类代理核实，因此会首先抽取最有可能相关的论文。

3. **随机**：随机 查询策略从未标注列表中随机抽取论文，不使用任何信息性度量。在这种情况下，不应用主动学习。

3.5 人机混合团队工作流程

我们假设人机混合团队的任务是筛选一组论文，以识别符合给定范围的论文。人工代理通过指定筛选标准来启动标题与摘要筛选过程。然后，人工智能代理随机抽取一部分论文，由人工代理进行筛选，作为初始训练集。在此基础上，我们的模型经过训练，从训练集中学习筛选标准。借助所学知识，人工智能代理预测未筛选论文的预测优先级分数。根据预测优先级分数，人工智能代理需要检查筛选—训练—预测—采样循环是否应当停止。如果不应停止，则采用某种策略抽取一组论文，供下一轮筛选。

人工智能代理决定在经过几轮迭代后停止训练模型时，会根据预测的 PS 值对所有尚未筛选的论文进行优先排序。随后，人类代理分批筛选优先排序的论文，并决定在哪一批结束筛选过程。由于数据集不平衡，因此对训练集采用随机过采样方法，使

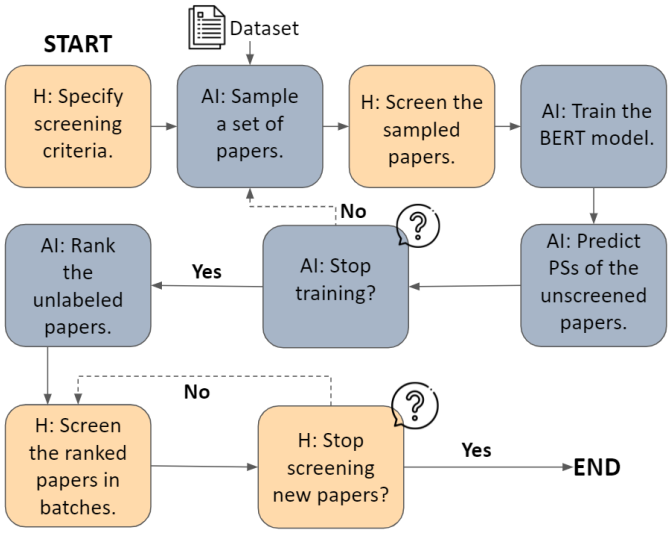


图 4：人机混合团队的工作流程。H 代表人工代理。

两个类别的样本数量相等。在本研究中，人工智能代理每次抽取 1,000 篇论文作为一批。选择这一批次规模，是因为它平衡了更新模型所带来的收益与成本。

3.6 评估指标

在机器学习领域，准确率和 F1分数通常用于评估分类模型的性能。然而，仅凭这些指标不足以评估人机混合团队的性能和效率。本研究根据定义 2 中的人工工作量 来评估人机混合团队的性能，即实现定义 3 中纳入率 所需的人工工作量。计算成本从另一个角度反映了团队效率，但本研究不讨论这一点。

定义 2。 人工工作量 定义为比率 (HE)：为识别一定数量的相关论文而需要人工筛选的论文数量 ($n_{screened}$) 与数据集中的论文总数 ($n = 68$，在模拟 EGM 设计案例研究中为 539 篇)之比，计算公式为：$HE = n_{screened}/n$。

定义 3. 纳入率 是已识别出的纳入论文数量（$n_{identified}$）与数据集中纳入论文总数（$n_{included} = 5$，在模拟的 EGM 设计案例研究中为 281）之比（IR），计算公式为：$IR = n_{identified}/n_{included}$。资源有限时，纳入率越高越好。

给定一组科学论文和筛选标准，高效的人机混合团队应在达到令人满意的纳入率的同时，尽可能减少人工工作量和计算成本；或者在现有人工工作量和计算资源的条件下，最大限度地提高纳入率。此外，我们还报告了每种情况下相应分类模型的 F1分数，以评估人工智能代理的性能。

4. 实验设置

本节讨论两个案例研究的实验设置：已部署的 EGM 设计和模拟 EGM 设计。每个案例研究特有的方面将在4.1和4.2节中介绍，而基线技术和实现细节等共同部分将在4.3 和4.4节中介绍。

4.1 已部署的 EGM 设计实验

人机团队的任务是设计三个待部署的 EGM：农业、营养和韧性 EGM。对于每个 EGM，人机交互以及相应的实验都聚焦于标题和摘要筛选过程。

三个 EGM 独立创建，并遵循相同的流程。它们最初的数据集规模如下：农业 221k、营养 117k、韧性 60k。每个 EGM 的标题和摘要筛选所采用的人机工作流程如图 4 所示，并已在第 3.5 节中介绍。每个 EGM 的流程完成后，人机团队会对各自数据集中的一个小型子集进行标注。我们将使用这些已标注数据集开展以下实验。

4.1.1 模型性能实验。 我们探究行业标准支持向量机方法以及我们提出的 BERT 方法在三个证据图谱中将文档分类为相关或不相关文档方面的有效性。由于本案例研究具有真实世界的性质，我们仅拥有所选择进行筛选的文档的标签。对于每个已标注数据集，我们执行 85% - 15% 的训练-测试集划分，并确定分类准确率。

4.1.2 人力投入实验。我们还从人工工作量的角度比较了训练后的 BERT 和支持向量机模型，以评估混合团队效率。具体而言，我们假设测试集中的文档将分别按照 BERT 和支持向量机模型预测的优先级评分降序进行筛选。人工工作量定义为：为达到特定纳入率，人类需要筛选的文档所占的百分比。如果要获得相同数量的纳入文档，所需筛选的文档越少，混合团队的效率就越高。也就是说，所需的人工工作量越少。

4.2 模拟环境治理机制设计实验

在本案例研究中，假设人类与人工智能组成的混合团队需要筛选一组共 68,539 份文档，以找出符合给定范围的文档。该数据集已完成全部标注，因此我们可以测试不同训练集规模和主动学习采样策略的效果。

4.2.1 训练规模实验。对于机器学习模型的训练，更大的训练集通常有助于提升模型性能，但需要投入更多人工工作量来标注数据。在人机混合团队中，应仔细权衡模型性能与标注所需的人工工作量，以实现较高的团队效率。我们开展实验，探究训练规模如何影响混合团队效率，具体考察模型性能（F1分数）和所需人工工作量。具体而言，我们以包含 1,000 篇论文的初始训练集为起点；为扩充训练集，在每次迭代中，我们随机抽取 1,000 篇论文，对其进行标注并将其加入训练集，迭代范围从

1,000 篇到 6,000 篇。在训练过程中，我们使用训练集中 85% 的论文来训练模型，并将 15% 用作验证集。其余所有论文构成测试集。

4.2.2 主动学习实验。 当人工智能代理抽取新的待筛选论文时，所使用的查询策略会影响抽取论文的信息量，进而影响机器学习模型性能和混合团队效率。我们通过实验将两种不同的采样策略——LC 和 HP——与随机抽样进行比较。对于每种采样策略，我们都从与随机抽样情形相同的初始训练集开始，该训练集包含 1,000 篇论文。之后，在每次迭代中，我们使用 LC 或 HP 策略抽取 1,000 篇新论文，以扩充训练集。我们针对这两种采样策略，测试了从 1,000 到 7,000 不等的训练集规模。

4.3 基线

为回答 RQ1——当人工智能代理在最优数据规模下接受训练时，能够节省多少人工工作量？，我们将不同训练集规模实验中的最佳结果与基线情况进行比较。在第一种情况下，人工团队独立完成相同的任务，不借助任何人工智能辅助。也就是说，人类代理从数据集中随机筛选论文。第二种情况采用基于支持向量机（SVM）的分类器，该分类器专为检索随机对照试验而开发，并已集成于 EPPI-Reviewer 软件中 [22]。对于第二种基线情况，我们还针对从 1,000 到 7,000 的四种不同训练集规模进行了实验，并将其中表现最佳的模型用作基线。

为回答 RQ2 - 通过主动学习增强混合团队，能够进一步节省多少人工工作量？，我们将不同训练集规模实验中表现最佳的模型作为基线，其中所有抽样论文均为随机选取。我们分别将 LC 抽样策略和 HP 抽样策略实验中表现最佳的模型与基线模型进行比较。

4.4 实现细节

在本研究中，我们以 1×10^{-5} 的学习率训练模型。训练过程开始时设有一个预热阶段，持续一个 epoch。实验在 Intel(R) Xeon(R) W-2295 CPU @ 3.00GHz 3.00 GHz 处理器上进行，该处理器具有 18 个核心和 256 GB 随机存取存储器。模型训练和预测均在 Nvidia RTX A5000 GPU 上进行（每次运行使用一块 GPU）。每项实验重复五次。使用 LC 和 HP 策略抽样获取新论文时，若需用到预测的不确定性和 PS，则采用五次运行所得预测的均值，以提高结果的可重复性。

5. 结果

在以下各节中，我们将展示两个案例研究的实验结果：已部署的 EGM 设计和模拟 EGM 设计。我们比较不同的机器学习模型、训练集规模和主动学习采样策略，并报告它们对模型性能和人工工作量的影响。此外，我们还将讨论本研究的局限性，以及人机团队未来在 EGM 设计中的应用。

5.1 已部署的 EGM 设计结果

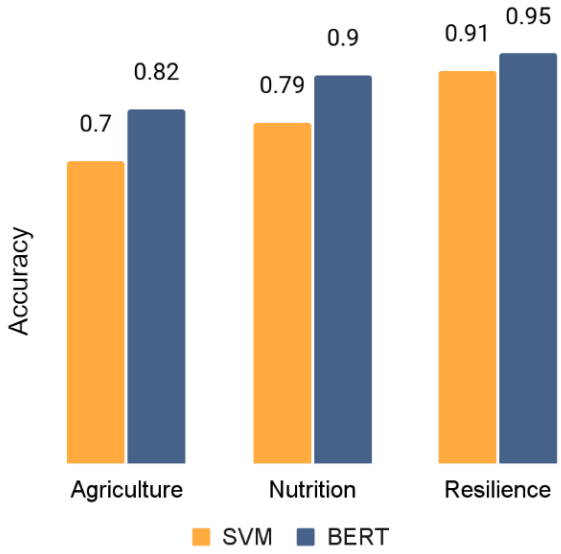


图 5：为创建的三个 EGM（农业、营养和韧性）构建的 BERT 和支持向量机模型的分类准确率。

本节比较由 BERT 和支持向量机模型完成的纳入分类。我们从两项不同的指标比较模型性能：总体准确率和节省的工作量。BERT 模型是我们提出的方法，而支持向量机模型则是 EPPI Reviewer 等工具用于文档分类的方法，因此代表了行业标准。

5.1.1 模型准确率。图5 展示了我们创建的三个 EGM 中，BERT 和支持向量机分类模型各自的准确率。结果表明，在所有三个 EGM 中，BERT 模型的准确率都高于支持向量机模型。目前用于辅助创建 EGM 的最常见自然语言处理工具基于支持向量机模型，因此我们的结果表明，使用 BERT 进行分类优于行业标准的 EGM 创建工具。

5.1.2 节省的工作量。图 6 展示了纳入率如何随人工工作量变化。橙色的“理想”线表示完美的纳入率，即只筛选相关文档，因此所有已查看的文档都会被纳入。灰色的“无机器学习”线表示这样一种情况：人类专家必须随机筛选所有文档，才能找到所有纳入的文档。我们比较了两种机器学习策略——BERT 和支持向量机，发现 BERT 在所有三个 EGM 中的表现都优于支持向量机。

我们对三个 EGM 的筛选论文进行了一系列实验，以作比较。为了使筛选论文的纳入率达到 80%，我们发现，与基线 EPPI-Reviewer 的支持向量机相比，使用基于 BERT 的模型时，人工评审员在农业领域所需的人工工作量减少了 47%，在营养领域减少了 17%，在韧性领域减少了 75%。三个 EGM 中 BERT 和支持向量机所需人工工作量的原始数值见表 2。在实际筛选过程中，BERT 模型节省人工工作量的能力进一步增强，其中这些模型

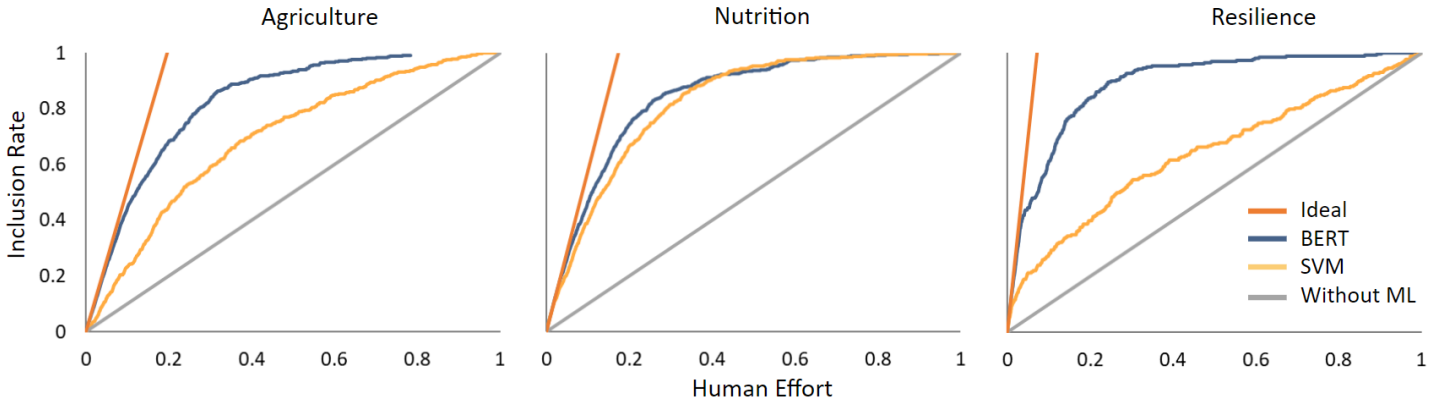


图 6：三个已部署 EGM 的纳入率随人工工作量的变化情况。以较低的人工工作量达到较高的纳入率更为理想。BERT 在三个 EGM 中的表现均优于支持向量机。

EGM	支持向量机 BERT		工作量百分比 BERT 节省的工作量
农业	53%	28%	**47%**
营养	29%	24%	**17%**
韧性	68%	17%	**75%**

表2：针对三个 EGM 和两种机器学习模型（支持向量机与 BERT），达到 80% 纳入率所需的人工工作量。BERT 节省的工作量百分比，按 BERT 与支持向量机所需人工工作量之间的百分比差计算。

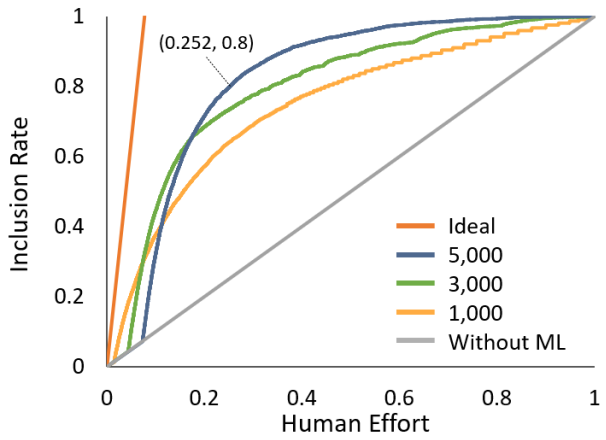


图7：当使用不同的训练集规模训练我们基于 BERT 的模型时，纳入率如何随人工工作量变化。训练规模为 5,000 时表现最佳，如图所示，它以最低的人工工作量达到了 0.8 的纳入率。

随着新的已标注文档作为训练数据不断加入，模型会多次更新。在这种情况下，模型会随时间不断迭代改进。随着分类准确率提高，模型可以仅向人工评审员推荐最相关的文档。这种主动学习方式将在模拟数据集中进行探讨，并在第5.2.2节中加以说明。

5.2 模拟EGM设计结果

5.2.1 节省工作量。我们通过纳入率（IR）和人工工作量（HE）评估人工智能人类混合团队的表现。图7 展示了模型在不同训练集规模下训练时，纳入率随人工工作量的变化情况。图中的灰色“无机器学习”线表示人工代理在没有任何人工智能代理协助的情况下独立工作的情形。由于人工代理会从数据集中随机筛选论文，因此此时纳入率等于人工工作量。靠近 y 轴的橙色“理想”线表示理想情形：每篇经过筛选的文档都是纳入的文档，且不筛选任何排除的文档。该线的斜率为 5,281（纳入论文总数）/ 68,539（数据集中的论文总数）。图中的其他曲线则说明，当基于BERT的人工智能代理在不同规模的数据集上接受训练时，人工代理投入更多筛选工作量后，纳入率如何变化。

每条曲线由两部分组成。第一段直线部分表示人工代理从数据集中标注论文以准备训练集的过程。由于筛选论文是随机选取的，因此纳入率等于 HE。训练集准备就绪后，我们使用该训练集训练模型，以预测未标注论文的 PS。直线之后的曲线部分对应人工代理根据预测优先级分数依次筛选未标注论文的过程。由于未标注论文会按优先级进行筛选，因此与第一部分相比，第二部分的曲线陡峭得多；第一部分的斜率与“Without ML”线相同。在曲线部分，初始斜率接近理想线的斜率，随后逐渐减小。这一趋势表明，与优先级分数较低的论文相比，优先级分数较高的论文更有可能被识别为纳入论文，这说明人工智能代理能够有效地确定未标注论文的筛选优先级。

由于高性能的人机混合团队能够以较低的 HE 实现较高的纳入率，其初始斜率应更接近7中的“理想”线。随着训练规模增大，曲线变得更陡，表明模型性能有所提升。这与图9中 F1 分数不断提高的趋势一致。然而，较大的训练规模需要更多的人力标注工作量（即沿对角线的第一部分中较长的直线），因此也可能降低人机混合团队的效率。

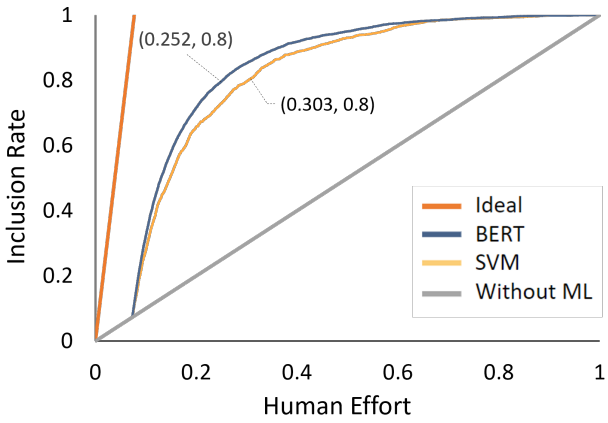


图8：不同机器学习模型的纳入率随人工工作量的变化情况：我们的模型（BERT）、行业标准（支持向量机），以及"理想"和"不使用机器学习"基线。

以 80% 的纳入率为目标时，曲线表明训练规模为 5,000 时，混合团队的效率最高。在这种情况下，人类代理只需筛选 25.2% 的论文即可达到 80% 的纳入率；而在没有人工智能指导的情况下，要达到相同的纳入率，他们需要筛选 80% 的论文。因此，将基于 BERT 的人工智能代理纳入人类团队后，达到 80% 纳入率所需的人工筛查工作量可减少 54.8%。

我们还比较了基于BERT的模型与EPPI-Reviewer软件中使用的支持向量机模型作为人工智能代理时的有效性。同样地，我们使用不同的训练集规模（1,000、3,000、5,000、7,000）训练支持向量机模型，其中训练集规模为5,000时，达到80%的纳入率所需的人工工作量最少。图8比较了效果最佳的基于BERT的模型（5,000）和效果最佳的支持向量机模型（5,000），结果表明，与支持向量机模型相比，基于BERT的模型使人工代理在达到任意纳入率时都能节省更多筛选工作量。具体而言，在达到80%的纳入率时，使用基于BERT的人工智能代理的人工代理比使用基于支持向量机的人工智能代理可多节省5.1%的筛选工作量。因此，我们基于BERT的模型作为人工智能代理时更加有效。

5.2.2 主动学习的影响。 在以下部分中，我们将讨论用于抽样新数据以扩大训练集规模的策略如何影响人工智能代理的性能，以及人机混合团队在TA筛选任务中的效率。

主动学习、训练集规模与模型性能

我们在此报告使用不同训练集规模和不同抽样策略开展的实验结果，以展示将人工智能代理纳入人类团队所产生的影响，并回答RQ1。按照不平衡数据集分类问题的实验方案，我们将以默认阈值0.5计算的F1分数作为分类指标。在这些实验中，抽样的论文均为随机选取。图9 中的黑色曲线显示了F1分数随训练集规模的变化。当训练集规模增加时，F1分数会逐渐提高，但边际效应递减，尤其是在训练集规模大于5,000之后。

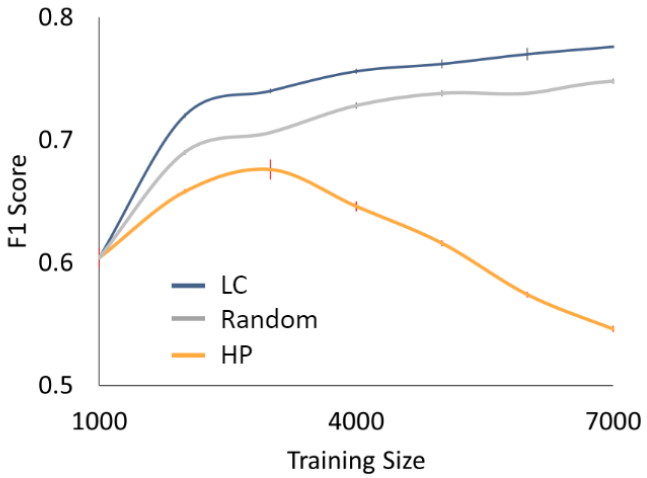


图9：模型性能（以F1分数表示）随训练集规模和主动学习抽样策略的变化。柱状图表示一个标准误。我们发现，最低置信度（LC）策略的表现最佳。

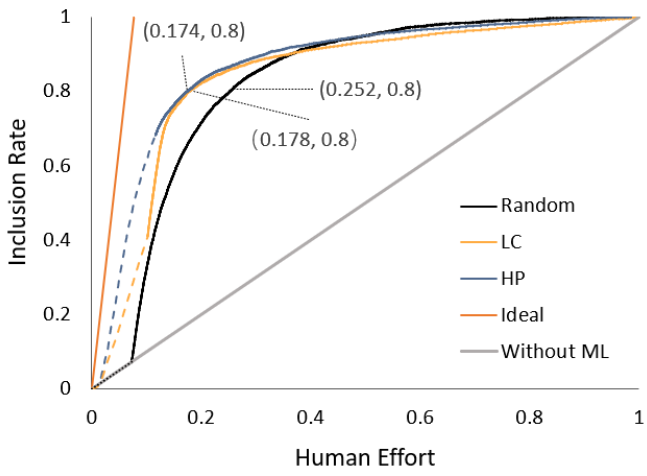


图 10：不同主动学习采样策略对人工工作量与纳入率关系的影响。每条曲线中的虚线部分表示筛选—更新—预测—采样迭代过程，而实线部分对应于人工代理筛选优先级排序后的论文的过程。

与随机抽样的情况类似，训练规模会影响我们的分类模型的性能。如图9所示，采用最低置信度策略时，增大训练规模会提升 F1 分数。若采用最高优先级策略，适中的训练规模（例如 2,000）对 F1分数最有利；当训练集规模超过某个值（例如 2,000）时，继续增大训练集会降低 F1分数。总体而言，使用最低置信度策略对新论文进行抽样，比随机抽取新论文能得到更好的分类模型，这一点从更高的 F1分数可以看出；然而，

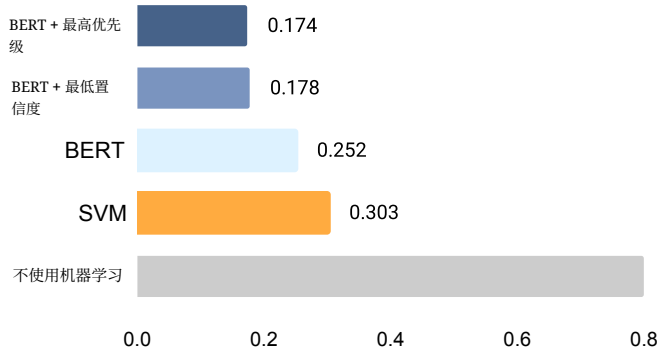


图 11：对于各种模型，达到 80% 纳入率所需的人工工作量。人工工作量越低越好。在不使用机器学习的情况下，需要投入 80% 的人工工作量才能达到 80% 的纳入率。

与随机抽样相比，最高优先级抽样策略得到的分类模型更差，F1 分数更低。

主动学习、训练规模与人工工作量

所选的主动学习抽样策略和训练规模也会影响人机团队的效率。在随机抽样条件下，适中的训练规模能够很好地平衡更高模型性能与创建训练数据所需更多标注工作量之间的权衡，从而实现最高的团队效率。主动学习也呈现出类似趋势。为达到 80% 的纳入率，分别采用最低置信度和最高优先级抽样策略时，训练规模为 7,000 的人机混合团队效率最高。

图10 比较了不同抽样条件下的团队效率，包括随机抽样、最低置信度抽样和最高优先级抽样。可以看出，采用主动学习后，人机混合团队的效率显著提高。采用最低置信度和最高优先级策略时，为达到 80% 的纳入率，人工代理分别可节省 7.4% 和 7.8% 的筛查工作量。具体而言，每条曲线中的虚线部分表示筛查—更新—预测—抽样迭代（即图4中的“人工智能：停止训练？”循环），而实线部分则对应人工代理根据最终人工智能模型的预测筛选优先排序论文的过程。

我们可以看到，LC 曲线和 HP 曲线的虚线部分比随机抽样曲线的虚线部分陡峭得多。这些趋势表明，与随机抽样策略相比，采用 LC 和 HP 抽样策略时，抽样论文中纳入的论文占比更高。也就是说，在筛选—更新—预测—抽样的迭代过程中，LC 和 HP 抽样策略，尤其是 HP，能够显著提高团队的筛选效率。这可以从抽样策略本身得到解释。LC 策略会抽取分类不确定性最高的论文。鉴于数据集高度不平衡，我们的模型对少数类论文的分类置信度较低，即，“1”类中纳入的论文，因此从少数类中抽取的论文更多。HP 策略会抽取预测优先级分数最高的论文，根据 PS 的定义，这些论文更有可能被纳入。

此外，通过对分类不确定性最高的论文进行采样，LC 抽样策略还使我们的模型能够比其他采样策略更有效地从人工标注中学习。这一点可以从图10 中观察到：蓝色曲线的实线部分在早期阶段比红色和黑色曲线的实线部分更陡；而且图9中 LC 的 F1 分数最高。

6. 讨论

图11 全面展示了我们比较的各种机器学习方法及其对人工工作量的影响。该图显示了在无机器学习辅助、采用基于支持向量机的模型、采用基于 BERT 的模型，以及采用结合 LC 或 HP 主动学习抽样策略的基于 BERT 的模型时，达到 80% 纳入率所需的人工工作量。我们发现，基于 BERT 的模型优于支持向量机，使人工工作量相对减少了 16.8%。结果表明，与不使用主动学习的 BERT 模型相比，主动学习策略可进一步将人工工作量减少约 30%。由于我们的目标是加快证据图设计流程，并降低该流程对资源和时间的要求，这一结果意义重大。

在混合团队中，人类代理与人工智能代理之间的有效互动和相互学习能够显著提升团队性能。采用 LC 抽样策略时，从人类代理流向人工智能代理的信息（即通过已标注论文传递的人类知识）以及从人工智能代理流向人类代理的信息（即通过抽样或优先排序的论文传递的人工智能预测）都能提高人机混合团队的效率。相比之下，采用 HP 抽样策略时，从人工智能代理流向人类代理的信息对提升混合团队效率起着主要作用，尤其是在筛查—更新—预测—抽样迭代期间。然而，在 HP 策略下，从人类代理流向人工智能代理的信息对提升模型性能的作用并不那么显著。

在实际的筛选过程中，我们只知道数据集中部分论文的标签，这意味着实际纳入率及其整体变化趋势都是未知的。在这种情况下，很难确定何时停止扩充训练集并更新人工智能代理，以及何时停止筛选优先排序的论文。未标注论文预测排名的变化幅度以及纳入率的增长速度可以为我们提供停止依据。论文排名变化较小以及纳入率增长速度较低，可能分别意味着我们可以停止更新人工智能代理和停止筛选优先排序的论文。

根据十二名参与农业发展证据图谱工作的人工筛选人员的记录，我们得知，人工筛选人员平均每小时可筛选 38.6 (SE = 1.00) 篇论文。据此，与不使用人工智能代理相比，经主动学习增强的人工智能代理可为人工筛选人员节省 (80% − 17.4%) ×68,539/38.6 = 1,111.5 小时的标题与摘要筛选时间。与使用 EPPT-Reviewer 支持向量机作为人工智能代理的情况相比，我们的模型可节省 (30.3% − 17.4%) ×68,539/38.6 = 229.1 小时。

6.1 关于人工智能辅助设计证据图谱的讨论

我们为美国国际开发署创建可供部署和使用的 EGM 的工作性质，意味着团队面临着许多现实挑战。本节将讨论在设计由人工智能辅助的 EGM 时出现的独特挑战和局限性。

沟通成本 在人机团队中，沟通成本可能相当高昂，但难以量化，因为它涉及多个因素。其中一个因素是通过电子邮件交换信息所需的时间和精力，这可能导致延误和潜在的沟通误解。此外，更新文档标签和合并数据集可能是一项复杂且耗时的任务，需要监督和项目管理，以确保准确率。这些活动也可能造成错误，对人工智能模型的性能产生负面影响。最后，人类为文档标注、人工智能代理接收文档并更新模型，以及随后由人工智能代理发回新排序的文档供人类筛选，这些环节之间存在时间差，意味着团队可能在数据不完整的情况下开展工作。因此，优化沟通渠道并实施高效的沟通规程，有助于降低人机团队协作的相关成本。

对人工智能的信任 许多人工智能推荐系统面临的一项重大挑战是“冷启动”问题。人工智能代理必须在第一次迭代时对文档作出一些预测，但此时模型对这个新领域一无所知。在我们的案例中，我们使用全球发展领域的文档对模型进行了预训练，但这并不能保证它在没有额外训练数据的情况下，能够很好地对特定于农业的证据图谱文档进行分类。这一挑战虽很常见，但如果人类团队发现初始排序有误，可能会因此不信任人工智能。此外，我们模型的训练数据是由人们提供的标签，这些标签可能存在噪声。人工智能模型的性能受训练数据质量的制约，因此，要获得有意义且准确的模型结果，我们必须从一致的高质量训练数据开始。

“何时停止”：一项业务决策 已部署的 EGM 设计案例研究中的另一项挑战，是确定何时停止筛选过程，这是图4中所示的第二个问题。我们的人机团队在此面临两难：是继续筛选更多论文以提升模型性能，还是停止筛选，进入 EGM 流程的下一步（图2）。归根结底，这是一项业务决策，团队必须权衡改进模型以及识别最“真正阳性”文档所需的资源成本。我们尝试了两种确定“何时停止”的方法。第一种方法是：按照优先级分数对文档进行排序，然后计算两次连续迭代之间文档排名的相似度。如果在一次筛选迭代后，两次排名的相似度高于某个数值，我们就可以停止更新 BERT 模型。第二种方法是在达到特定的实时纳入率时终止人工筛选过程，例如，在当前迭代中筛选 1,000 份文档后识别出的相关文档数量。如果相关文档数量低于给定阈值，人工筛选团队就可以停止筛选过程。未来的工作可以专门研究何时停止筛选，因为这是人机团队需要作出的一个高度相关的决策。

全文筛选的自动化 如图所示，证据图谱设计流程 2包括标题和摘要筛选以及全文筛选。后续工作可以采用自然语言处理来辅助全文筛选步骤。然而，这一步在获取全文文档方面存在实际操作上的挑战。虽然许多机构订阅了学术文章数据库，但在不同机构内部及机构之间开展工作时，版权问题使全文文档的下载和使用颇具挑战。这最终决定了我们的项目范围仅限于标题和摘要筛选过程。

此外，要进行全文筛选，还需要训练另一个 BERT 模型，对全文文档进行纳入分类。这意味着人工筛选人员需要为这项任务生成训练数据集，这需要投入大量人工工作量。大语言模型（LLM）或许能应对这一挑战。经过数十亿份文档训练的大语言模型 [42]对语言有广泛的理解，未来的工作可以探讨它们能否在没有领域专用训练数据的情况下对全文文档进行分类。

反事实分析 我们的团队面临一项挑战：如何在真实场景中准确比较不同的文档筛选技术，例如使用基于 BERT 的模型、基于支持向量机的模型，以及不使用机器学习模型。人工评审员无法为每种技术分别创建三次证据图谱（EGM），以比较每种技术的完整流程。因此，在人机团队标注了部分数据后，我们通过开展回顾性实验来统一比较方式。我们使用这一已标注的子集呈现结果。我们还通过纳入第二项案例研究——模拟 EGM 设计，进一步尝试解决这一局限。在这项案例研究中，我们使用了一个包含 68,539 份文档的完整标注数据集，以便针对各种机器学习模型和主动学习采样策略开展实验。然而，这项挑战意味着，我们无法对在没有任何人工智能辅助的情况下 EGM 流程会如何推进进行真正的反事实分析。未来的工作可以针对每种不同的技术多次创建每个 EGM，以进一步解决这一局限。

6.2 人机团队未来在EGM设计中的应用

人工智能的自然语言处理子领域正在快速发展。OpenAI 的 ChatGPT [42], 和 Meta 的 Galactica [43]等大语言模型，正在改变世界对预训练语言模型的认知、利用方式及交互方式。这些模型在我们完成证据图谱创建后才发布；不过，我们预计它们的能力将改变自然语言处理在证据图谱设计中的应用方式。

我们进行了一系列探索性实验，以了解大语言模型在证据图谱设计方面的能力。我们通过提问“农业转型如何改变贫困、移民和粮食安全？”来探究 ChatGPT 对某些干预与结果之间关系的理解。大语言模型捕捉到了干预（农业转型）与结果（贫困、移民和粮食安全）之间的一般定性关系，但没有给出任何定量

影响、潜在信息来源，或有关这些关系得到何种程度研究的说明。这表明，ChatGPT 能够以较粗略的分辨率捕捉干预-结果关系，但无法提供人类团队或人机混合团队能够捕捉到的所有详细信息。

在一项小规模研究中，我们探讨了大语言模型能否识别图1所示证据图谱中十份文档的干预措施和结果。我们向 ChatGPT 提供了相关摘要，并要求它指出每篇摘要中的干预措施和结果。我们发现，大语言模型能够生成相关的干预措施和结果。其中一个回答是：“干预措施是一项农业转型计划，论文中未作进一步说明。研究评估的结果变量包括贫困、迁移、粮食安全和农业收入。”这种生成能力在证据差距图设计的早期阶段可能很有用；在这一阶段，人类专家会确定界定证据图谱范围的干预措施和结果类别（如图1中的列标题和行标题所示）。生成式文本也有助于为大量文档创建简短摘要，这符合证据综合的总体目标。

7. 结论

本文研究了：(1) 将基于BERT的人工智能代理纳入人工团队后，如何影响证据差距图设计流程中的团队效率；(2) 通过主动学习增强混合团队，如何提高混合团队效率。我们提出了一种应用于证据差距图设计流程中标题与摘要筛选阶段的人机混合协作工作流。我们 a) 为全球发展领域的农业、营养和韧性三个主题设计并部署了三个证据差距图，b) 使用一个完整标注的数据集开展模拟实验，以回答上述研究问题。结果表明，用于训练人工智能代理的数据规模会影响混合团队效率。当训练规模经过优化时，与不使用人工智能辅助相比，纳入基于BERT的人工智能代理可使人工工作量减少68.5%；与使用基于支持向量机的代理相比，在达到80%的纳入率时可使人工工作量减少16.8%。此外，通过主动学习增强混合团队，与不使用主动学习的BERT方案相比，可进一步减少30%的人工工作量。所提出的人机混合协作工作流已在三个证据差距图的实际构建过程中得到验证。因此，经过主动学习增强的人机混合团队能够显著加快全球发展领域的证据差距图设计和决策。

致谢

本出版物得以完成，离不开美国国际开发署韧性与粮食安全局（美国国际开发署）的资助支持。资助方未参与研究设计、数据收集与分析、发表决定或手稿撰写。本文所表达的观点仅代表作者个人立场，并不一定反映美国国际开发署的观点。

参考文献

[1] 美国国际开发署。加强循证发展。美国国际开发署（2016年）。[2] 115届国会，美国。《2018年循证政策制定基础法》。（2018年）。[3]Kraybill, David 和 Mercier, Stephanie。《美国如何从对发展中国家的农业和粮食安全投资中获益》。（2019年）。网址 https://www.usaid.gov/sites/default/files/documents/1867/BRIEF_-_US_Benefits_Overview.pdf。[4] White, Karen。《出版物产出：美国趋势与国际比较》。https://ncses.nsf.gov/pubs/ nsb20206/（2019年）。访问日期：2022年1月14日。[5] Donnelly, Christl A., Boyd, Ian, Campbell, Philip, Craig, Claire, Vallance, Patrick, Walport, Mark, Whitty, Christopher J. M., Woods, Emma 和 Wormald, Chris。《让证据综合更有助于政策制定的四项原则》。《自然》第558卷（2018年）：第361–364页。[6]Snilstveit, Birte, Vojtkova, Martina, Bhavsar, Ami, Stevenson, Je nnifer 和 Gaarder, Marie。《证据与空白地图：促进循证政策和战略研究议程的工具》。《临床流行病学杂志》第79卷（2016年）：第120–129页。DOI 10.1016/j.jclinepi.2016.05.015。访问日期：2019年7月6日，网址 https://www.jclinepi.com/article/S0895-4356(16)30190-1/abstract。[7] 3ie。《证据缺口图》。可见于 https://www.3ieimpact.org/evidence-hub/evidence-gap-maps（2021/08/04）（2021年）。[8]NLPIE。《BioMedICUS》。https://nlpie.github.io/ biomedicus/（2019年）。[9] Bommarito, Michael J, Katz, Daniel Martin 和 Detterman, Eric M.。《LexNLP：用于法律和监管文本的自然语言处理与信息提取》。（2018年）。网址 1806.03688。[10] Porciello, Jaron, Ivanina, Maryia, Islam, Maidul, Einarson, Stefan 和 Hirsh, Haym。《利用机器学习加速可持续发展目标的循证决策》。《自然·机器智能》第2卷（2020年）：第559–565页。[11] Briner, Rob B. 和 Denyer, David。《系统综述和证据综合作为实践与学术工具》。《循证管理手册：公司、课堂与研究》。纽约大学出版社（2012年）：第112–129页。[12] Blaizot, Aymeric, Veettil, Sajesh K., Saidoung, Pantakarn, Moreno-Garcia, Carlos Francisco, Wiratunga, Nirmalie, Aceves-Martins, Magaly, Lai, Nai Ming 和 Chaiyakunapruk, Nathorn。《在健康科学系统综述中使用人工智能方法：一项系统综述》。《研究综合方法》第13卷第3期（2022年）：第353–362页。DOI https://doi.org/10.1002/jrsm.1553。网址 https://onlinelibrary.wiley.com/doi/pdf/10.1002/jrsm.1553，网址 https://onlinelibrary.wiley.com/doi/abs/10.1002/jrsm.1553。[13] Altena, Allard, Spijker, René 和 Olabarriaga, Silvia。《系统综述中自动化工具的使用》。《研究综合方法》第10卷（2018年）。DOI 10.1002/jrsm.1335。

[14] 3ie。《证据映射》。可见于 https: // www.3ieimpact.org/evidence-hub/evidence-gap-maps（ 2021/08/04）（2021年）。[15] O'Mara-Eves, Alison, Thomas, James, McNaught, John, Miwa, Makoto 和 Ananiadou, Sophia。《在系统综述中使用文本挖掘进行研究识别：当前方法的系统综述》。《系统综述》第4卷第1期（2015年）：第1页。访问日期：2016年3月11日，网址 http://systematicreviewsjournal.biomedcentral. com/articles/10.1186/2046-4053-4-5。00000。[16] Oliver, Kathryn, Innvar, Simon, Lorenc, Theo, Woodman, Jenny 和 Thomas, James。《政策制定者使用证据的障碍与促进因素：一项系统综述》。《BMC卫生服务研究》第14卷（2014年）：第2页。DOI 10/gbfqbx。访问日期：2016年7月30日，网址 http://dx.doi.org/ 10.1186/1472-6963-14-2。00000。[17] Mikolov, Tomas, Sutskever, Ilya, Chen, Kai, Corrado, Greg 和 Dean, Jeffrey。《词与短语的分布式表示及其组合性》。（ 2013年）。网址 1310.4546。[18] Mikolov, Tomas, Chen, Kai, Corrado, Greg 和 Dean, Jeffrey。《在向量空间中高效估计词表示》。（2013年）。网址 1301.3781。[19] Pennington, Jeffrey, Socher, Richard 和 Manning, Christopher D.。《 GloVe：用于词表示的全局向量》。EMNLP，第14卷：第1532–1543页。2014年。[20] Ouzzani, Mourad, Hammady, Hossam, Fedorowicz, Zbys 和 Elmagarmid, Ahmed。《 Rayyan——一款用于系统综述的网页和移动应用》。《系统综述》第5卷第210期（2016年）。DOI https://doi.org/10.1186/s13643-016- 0384-4。网址 https://link.springer.com/article/10.1186/ s13643-016-0384-4#citeas。[21] Marshall, Iain J, Kuiper, Joel, Banner, Edwards 和 Wallace, Byron C.。《自动化生物医学证据综合：RobotReviewer》。2017年。DOI 10.18653/v1/P17-4002。[22]Thomas, James, McDonald, Steve, Noel-Storr, Anna, Shemilt, Ian, Elliott, Julian, Mavergames, Chris 和 Marshall, Iain J.。《机器学习在漏掉研究的风险极低的情况下减少了工作量：随机对照试验分类器的开发与评估》。《临床流行病学杂志》第133卷（2021年）：第140–151页。DOI https://doi.org/10.1016/j.jclinepi.2020.11.003。网址 https://www.sciencedirect.com/science/article/pii/ S0895435620311720。[23] Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N., Kaiser, Lukasz 和 Polosukhin, Illia。《 Attention Is All You Need》。（2017年）。网址 1706.03762。[24]Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton 和 Toutanova, Kristina。《BERT：深度双向 Transformer 的预训练及语言理解》。（2019年）。网址 1810.04805。[25] Liu, Yinhan, Ott, Myle, Goyal, Naman, Du, Jingfei, Joshi, Mandar, Chen, Danqi, Levy, Omer, Lewis, Mike, Zettlemoyer, Luke 和 Stoyanov, Veselin。《RoBERTa：一款 Ro

《高效优化的 BERT 预训练方法》。（2019年）。URL 1907.11692。[26] Radford，Alec、Wu，Jeffrey、Child，Rewon、Luan，David、Amodei，Dario 和 Sutskever，Ilya。《语言模型是无监督的多任务学习者》。OpenAI 博客（2019年）。[27]Aum，Sungmin 和 Choe，Seon。《srBERT：使用 BERT 进行系统综述的自动文章分类模型》。Systematic Reviews，第 10 卷，第 285 期（ 2021 年）。DOI h t tps://doi.org/10.1186/s13643-021-01763- w10.1002/jrsm.1335。[28] Ein-Dor，Liat、Halfon，Alon、Gera，Ariel、Shnarch，Eyal、Dankin，Lena、Choshen，Leshem、Danilevsky，Marina、Aharonov，Ranit、Katz，Yoav 和 Slonim，Noam。《BERT 主动学习：一项实证研究》。EMNLP。2020年。[29]Aggarwal，Umang、Popescu，Adrian 和 Hudelot，Celine。《面向少数类的不平衡数据集主动学习》。2020年第 25 届国际模式识别会议（ICPR）。2021年。IEEE。DOI 10.1109/icpr48806.2021.9412182。URL https://doi.org/10. 1109%2Ficpr48806.2021.9412182。[30] Gal，Yarin、Islam，Riashat 和 Ghahramani，Zoubin。《利用图像数据进行深度贝叶斯主动学习》。（2017年）。DOI 10.48550/ARXIV.1703.02910。URL https://arxiv.org/abs/ 1703.02910。[31] Tur，Gokhan、Hakkani-Tur，Dilek 和 Schapire，Robert E.《结合主动学习与半监督学习进行口语理解》。Speech Communication，第 45 卷（2005年）：第 171–186 页。[32]Chung，Mu-Huan、Chignell，Mark、Wang，Lu、Jovicic，Alexandra 和 Raman，Abhay。《用于数据窃取检测的交互式机器学习：结合人类专家知识的主动学习》。IEEE 国际系统、人类与控制论会议（SMC）论文集：第 280–287 页。2020年。DOI 10.1109/SMC42975.2020.9282831。[33] Olsson，Fredrik。《自然语言处理背景下的主动机器学习文献综述》。技术报告，瑞典计算机科学研究所。2009年。[34] Angluin，Dana。《查询与概念学习》。Machine Learning，第 2 卷（1998年）：第 319–342 页。[35] Atlas，Les、Cohn，David 和 Ladner，Richard。《利用查询和选择性采样训练连接主义网络》。Touretzky，D.（编）。《神经信息处理系统进展》，第 2 卷。1989年。Morgan-Kaufmann。URL https://proceedings.neurips.cc/paper/ 1989/file/b1a59b315fc9a3002ce38bbe070ec3f5-Paper.pdf。[36] Lewis，David D. 和 Gale，William A.《训练文本分类器的序列算法》。第 17 届 ACM SIGIR 信息检索研究与发展会议年度国际会议论文集：第 3–12 页。1994年。Springer-Verlag，柏林，海德堡。[37] Settles，Burr。《主动学习文献综述》。计算机科学技术报告 1648。威斯康星大学麦迪逊分校。2009年。[38] Culotta，Aron 和 Andrew，McCallum。《减少结构化预测任务的标注工作量》。AAAI：第 746–751 页。2005年。

[39] Scheffer，Tobias、Decomain，Christian 和 Wrobel，Stefan。《用于信息抽取的主动隐马尔可夫模型》*IDA*。2001年。[40] Shannon，C. E.《通信的数学理论》《贝尔系统技术期刊》第 27 卷，第 4 期（1948年）：第 623–656 页。DOI10.1002/j.1538-7305.1948.tb00917.x。[41] 3ie。《发展证据门户》。访问地址：https://developmentevidence.3ieimpact.org （2021/08/04）（2021年）。

[42] OpenAI，TB.《ChatGPT：优化用于对话的语言模型。》*OpenAI*（2022）。[43] Taylor、Ross、Kardas、Marcin、Cucurull、Guillem、Scialom、Thomas、Hartshorn、Anthony、Saravia、Elvis、Poulton、Andrew、Kerkez、Viktor 和 Stojnic、Robert。《Galactica：面向科学的大型语言模型。》（2022）。DOI10.48550/ARXIV.2211.09085。网址https://arxiv.org/abs/2211.09085。