Relatedly：利用现有相关工作章节为文献综述搭建脚手架

Srishti Palani*
srishti@ucsd.edu 加州大学圣迭戈分校，美国

Aakanksha Naik
aakankshan@allenai.org 艾伦人工智能研究所，美国华盛顿州西雅图

Doug Downey
dougd@allenai.org
艾伦人工智能研究所，
美国华盛顿州西雅图

Amy X. Zhang
axz@cs.uw.edu 华盛顿大学，美国华盛顿州西雅图

Jonathan Bragg
jbragg@allenai.org
艾伦人工智能研究所，
美国华盛顿州西雅图

Joseph Chee Chang
josephc@allenai.org 艾伦人工智能研究所 西雅图，
华盛顿州，美国

arXiv:2302.06754v1 [cs.HC] 13 Feb 2023

摘要
研究科学主题的学者必须花时间阅读、提取含义，并在众多论文之间找出联系。随着科学文献不断增长，这项工作变得愈发困难。与此同时，作者会在论文的相关工作部分总结先前研究，但其范围仅限于支持单篇论文。一项形成性研究发现，阅读多个相关工作段落有助于概览某个主题，但要理清相互重叠和分歧的参考文献及研究重点并不容易。在这项工作中，我们设计了一个系统 *Relatedly*，为探索和阅读围绕某一主题的多个相关工作段落提供辅助。其功能包括动态重新排序和高亮显示，以突出尚未探索的不同信息；自动生成描述性段落标题；以及弱化冗余信息。通过一项被试内用户研究（*n=15*），我们发现，与基线论文列表相比，学者使用 *Relatedly* 能够生成更连贯、更有见地且更全面的主题大纲。

CCS 概念
• **以人为中心的计算** →**图形用户界面**。

关键词
文献综述、科学发现、探索式搜索、意义构建

ACM 参考文献格式：
Srishti Palani、Aakanksha Naik、Doug Downey、Amy X. Zhang、Jonathan Bragg 和 Joseph Chee Chang。2023。Relatedly：利用现有相关工作章节为文献综述搭建脚手架。载于*2023*年 *CHI*计算机系统中的人为因素会议论文集（*CHI '23*），*2023*年*4*月*23*日至*28*日，德国汉堡。ACM，美国纽约州纽约市，20页。https://doi.org/10.1145/3544548.3580841

*在艾伦人工智能研究所的 Semantic Scholar Research 研究实习期间完成。

特此授权免费制作本作品的部分或全部内容的数字或纸质副本，供个人或课堂使用，但不得出于营利或商业目的制作或分发副本，且副本首页必须附有本声明及完整引文。本作品中第三方组件的版权必须得到尊重。其他用途请联系所有者/作者。*CHI '23*，*2023*年*4*月*23*日至*28*日，德国汉堡© 2023 版权所有归所有者/作者所有。ACM ISBN 978-1-4503-9421-5/23/04。
https://doi.org/10.1145/3544548.3580841

1 引言
科学发现和创新依赖于学者对先前工作有深入的了解，而他们通过阅读文献、提取意义，并从大量含糊不清的领域特定信息中辨识众多论文之间的联系来实现这一点 [39, 74]。随着科学出版物的指数级增长 [5,20, 32, 71] 以及科学日益增强的跨学科性质[52, 72]，这一过程正变得愈发困难。遗憾的是，当前的方法（例如阅读综述论文或使用文本或可视化搜索引擎）在意义建构支持或时效性方面存在局限。例如，综述论文会对某个研究主题进行概述，呈现连贯的研究主题，并提供经过精心综合的描述 [6, 39]。但由于编撰综述论文需要付出大量人工努力，因此并非所有主题都有相应的综述论文；而且随着新研究不断涌现，综述论文也很快会过时。为解决这一问题，学者还经常依赖自动化方法来探索文献，例如使用学术搜索引擎，包括 Google Scholar[1] 和 Semantic Scholar[2]。这些工具有助于查找与查询相关的论文，却无法呈现连接多篇论文的更高层次主题。其他工具则利用可视化，根据引文或语义嵌入向量等指标将论文联系起来并进行聚类 [38]，例如 Connected Papers。[3] 然而，用户可能难以理解复杂图和聚类背后的含义，因为自动聚类往往会混淆多个维度 [25]。因此，当没有及时可用的综述论文时，学者仍需查阅许多单篇论文，并尝试找出其中的潜在主题及彼此之间的联系，以完成文献综述 [57]。

与此同时，学术论文的作者在需要撰写论文的相关工作章节时，也会经历类似的过程，探索并总结先前研究。相关工作部分提供了最新且经过充分综合的先前工作摘要 [68, 69], , 但由于其旨在支持单篇论文，因此往往无法像综述论文那样全面概述该主题。不过，如果读者能够看到来自不同论文、讨论同一主题的多个 相关工作章节，那么这一问题或许可以得到缓解，使他们

[1]谷歌学术[2]
www.semanticscholar.org[3]
www.connectedpapers.com

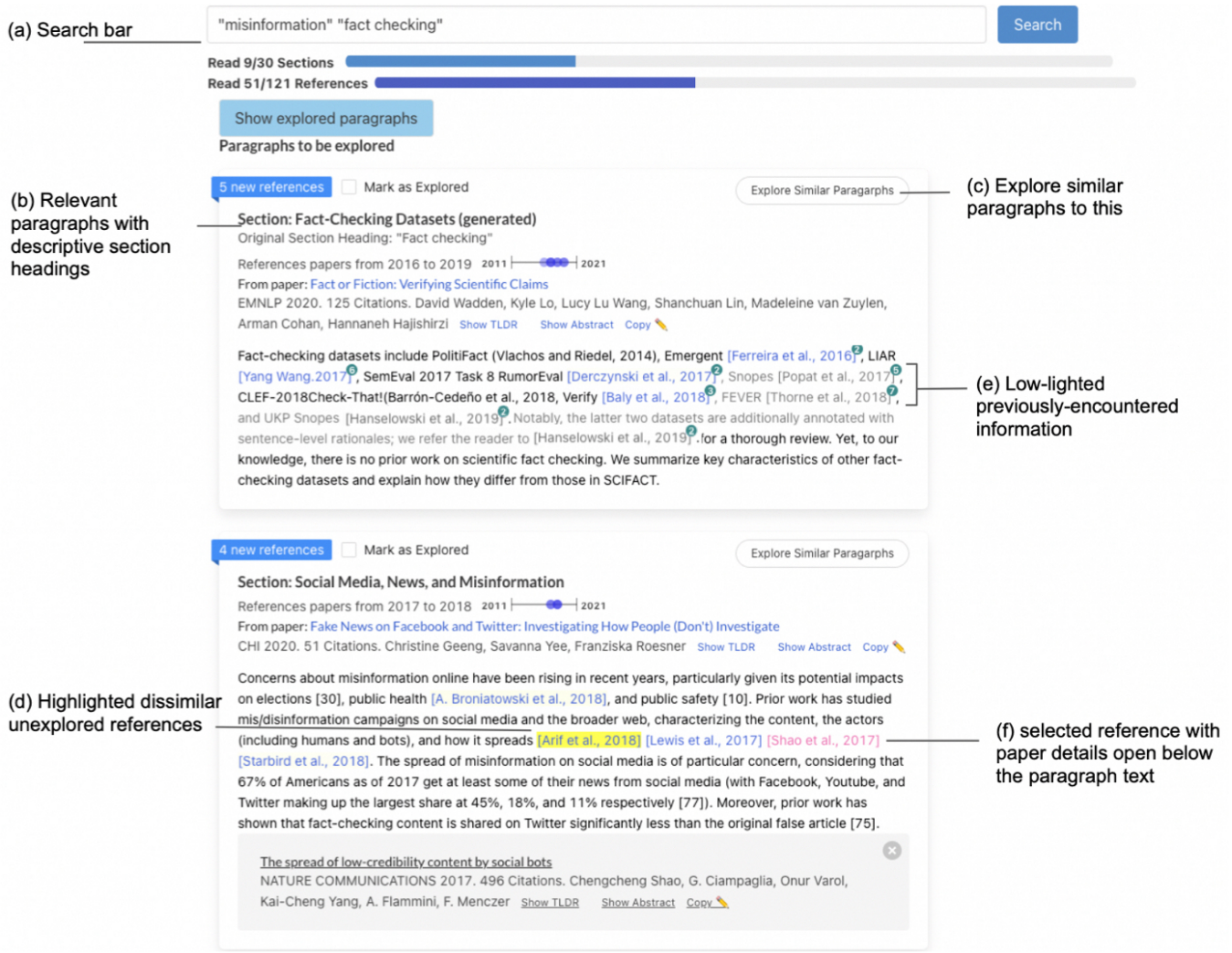

图1：*Relatedly*系统向用户展示某一主题的先前工作中的相关工作段落，并通过提供阅读、优先级排序和进度跟踪功能，为段落探索体验提供支持。此处，概览视图展示与高层次查询主题相关的段落，并按多样性排序，使排名靠前的结果涵盖广泛的子主题。对了解某个段落的子主题感兴趣的用户可以点击"探索相似段落"按钮，进入图2中的相似段落视图。

能够从多位作者那里获得对该领域更广泛的覆盖和不同的观点。

为进一步探究这一机遇，我们首先基于从众多论文中提取的一组相关工作章节构建了一个文本搜索引擎，并利用它对10位学者开展了形成性访谈研究。我们询问了学者，与他们当前的实践相比，他们在探索相关工作章节时的反应以及所面临的挑战。我们发现，尽管参与者更喜欢阅读相关工作章节，但由于不同的相关工作章节既有重叠的参考文献和关注点，也有分歧的参考文献和关注点，他们难以确定阅读优先级并跟踪阅读进度。

受访谈中的见解启发，我们设计了 *Relatedly*，这是一个用于支架式探索文献的新型系统，借助相关工作章节综合呈现广泛主题。如图 1 所示，当用户在 Relatedly 中查询某个主题（a）时，系统会从不同论文的相关工作部分中检索相关段落及其章节标题（b），帮助用户快速了解不同的研究脉络。如果段落没有描述性章节标题，Relatedly 会自动生成一个。用户还可以通过探索某个段落的相似段落，进一步深入了解某个子主题（c，接下来的图 2）。为了帮助用户安排阅读优先级并跟踪阅读进度，当用户在 Relatedly 中探索相关工作章节时，系统会记录用户读过的段落和参考文献，然后动态调整其余段落的排序，并突出显示与用户阅读历史不同的未探索参考文献（d）。

Relatedly 还会淡化提及已读段落中引用过的论文的句子（e）。

我们开展了一项被试内研究（$n = 15$）来评估 Relatedly。在研究中，参与者需要探索两个科学主题的文献，最终目标是分别为每个主题撰写一份综述论文提纲：一种条件下使用 Relatedly，另一种条件下使用返回论文列表的基线系统。结果发现，与基线条件相比，参与者使用 Relatedly 撰写的提纲质量更高；对条件不知情的主题专家对提纲进行了评估。系统日志显示，尽管每种条件下参与者的时间相同，且都能访问同一组论文，但使用 Relatedly 的用户接触的信息显著更多（包括段落和论文）。参与者还自我报告称，相比于浏览论文列表来开展文献综述，他们更喜欢使用 Relatedly 探索相关工作章节。

总而言之，本研究作出了以下贡献：

- 一种通过阅读和探索从多篇论文中提取的相关工作章节，来发现并系统综述科学主题文献的新方法。
- 形成性用户研究（$n = 10$）的结果，概述了当前的文献综述实践，以及用户在采用此方法时面临的挑战。
- *Relatedly* 系统通过阅读、优先级排序和进度跟踪功能，为相关工作段落的探索提供支架。
- 一项针对 15 名参与者的被试内研究所得的实证见解表明，支架式探索相关工作章节有助于发现和综合文献。

2 相关工作

我们的工作建立在先前工作之上，这些工作研究了科学家如何探索和综述文献，以及为支持这些复杂探索过程而构建的工具。

2.1 学者如何进行文献综述

文献综述有助于学者识别先前研究中的模式与空白，从而发现机会、确定开展新研究的理由，并将研究目标置于文献脉络之中 [68]。综述会详细介绍该主题中已知的研究成果以及尚待探讨的研究问题。高质量的文献综述会全面涵盖所选研究主题中的所有主要主题与子主题，既包括经典的奠基性研究，也包括近期研究，以展现对当前主题的深入理解 [16, 30, 39]。为实现这些目标，学者必须花时间全面探索某个主题，并阅读许多单篇论文。然而，试图通过阅读单篇论文来掌握某个领域全貌的意义建构过程，既耗时，也会造成巨大的认知负担 [6, 62, 69]。例如，用户可能难以拓宽阅读范围，以便快速识别不同的研究脉络。学者需要梳理的大量单篇论文和冗余信息，往往会导致信息过载 [51]。

学者应对这一问题的一种方式，是针对不同的研究主题撰写综述论文 [16, 30, 39]。然而，随着出版速度呈指数级增长，综述论文往往无法及时获得 [5, 20, 32, 71], ，即使能够获得，也会随着更新的研究不断涌现而很快过时。

与此同时，在大多数科学论文中，作者会在相关工作章节中总结多篇论文并梳理它们之间的联系，以便将自己的研究置于相关背景之中 [68, 69]。这些相关工作章节中的每个段落，都会为其中引用的单篇论文补充背景和结构。例如，一篇关于错误信息的论文，其相关工作部分可能会将一组被引用的论文归入题为“错误信息如何影响公共卫生”的段落，而另一组论文则可能归入“错误信息如何在社交媒体上传播”这一标题下。然而，相关工作章节只关注论文的特定视角，并不会试图全面概述更广泛主题中的所有主题和子主题。例如，上述关于错误信息的论文可能会将其相关工作部分聚焦于社交媒体上的健康错误信息，因为这与论文相关，但不会涵盖其他与错误信息有关的研究，例如检测错误信息的计算技术。因此，希望更全面了解某一主题相关文献的学者，可能需要阅读多篇论文中的多个相关工作章节。Relatedly 系统正是试图为这项任务提供支架。

信息觅食理论 [56] 为如何完成这一任务提供了一些指导。在复杂的探索性任务中，人们会在探索不同的信息斑块和利用已发现的信息斑块之间切换，以优化信息增益。他们依靠信息环境中的各种线索，即“信息气味”，来评估某个来源是否有助于获取信息。我们从信息觅食理论中汲取灵感，在 Relatedly 中提供信息气味线索，例如显示用户阅读每个段落可以了解到多少新信息。此外，为了支持从探索转向利用，Relatedly 允许用户根据某个段落深入查看相似段落；这使他们能够从不同观点出发，加深对某个子主题的理解。

2.2 支持文献综述的工具

如今，学者进行文献综述时最常依赖的工具之一是学术搜索引擎 [69], ，例如 Google Scholar[1] 和 Semantic Scholar[2]。这些工具在帮助用户查找与查询相关的单篇论文方面非常有效。然而，为了更深入地理解某个研究领域，例如开展文献综述时，学者通常需要综合多篇论文中的信息。理解论文之间的联系并揭示更大主题中不同且细微的研究主题，这一费力且耗时的过程在很大程度上需要用户自行完成，仅得到极少支持 [47, 63]。例如，在浏览搜索结果列表中的论文时，用户可能难以确定阅读优先级、跟踪散布在多篇论文中的信息，或了解自己在陌生信息空间中的总体进展。

分面搜索界面允许用户通过应用跨多个类别的筛选条件来浏览搜索结果 [25]。分类能够提供连贯且基本完整的标签。然而，人工分类耗时费力，而且难以持续更新。自动分类通常基于元数据 [25]。例如，Google Scholar 支持按时间筛选论文搜索结果

出版物和相关性等条件。类似地，Semantic Scholar 将'研究领域'、'出版物类型'等作为筛选维度，用户可以据此筛选论文。但元数据并非总是可用。此外，这些标签往往过于笼统，无法让人深入了解主题或领域。

论文可视化聚类试图采用另一种方法，帮助学者发现论文之间的关系。例如，给定一篇种子论文，Connected Papers[3] 利用引文图查找其他相关论文的聚类。Paper Quest[58] 和 Apolo [12] 等研究系统将一组作为输入的论文之间的引文关系进行可视化，并通过根据来源论文及其引文逐步展示更多相关论文，帮助用户克服信息过载。然而，关于聚类搜索界面的先前研究也指出，自动生成的聚类可能缺乏连贯性且难以理解，因为它们往往混淆多个维度 [25]。具体而言，论文可视化聚类方法通常会在相似论文之间显示边，但不会描述它们的语义关系 [25]。它们还会展示相似论文的聚类，却缺乏对其潜在主题的高层次描述 [25]。因此，学者仍需查阅单篇论文，以确定每个自动生成的聚类的含义，以及不同论文之间的关联方式[12]。

自动摘要技术，如多文档摘要 [17] 和科学地铁图 [64]，为原本复杂、难以理解的引文图添加了解释。然而，这些解释并不总是准确、连贯或全面。另一方面，人工系统（如 Threddy [33]）和众包系统（如 Knowledge Accelerator [10, 23, 35],、Crowdlines [46]）有助于为主题空间提供更连贯、全面且准确的摘要，但生成这些摘要需要时间和努力程度。

Relatedly 通过利用相关工作作者为论文构建高质量关联和聚类时已经付出的巨大努力，避开了生成这些内容这一难题。于是，主要挑战变成了如何探索多篇论文相互重叠的聚类，以及它们对论文之间关联方式的不同观点。

3 形成性研究与设计目标

为了了解用户在探索某个主题的相关工作段落并理解其内容时面临的挑战和采用的策略，我们与学者开展了一项形成性访谈研究。

3.1 形成性用户研究方法

我们对 10 位在搜索、阅读和撰写科学文献方面有三年以上经验的人进行了半结构化访谈（5 名男性、5 名女性，平均年龄 27.5 岁）。其中一人已获得博士学位，五人已获得硕士学位，四人已获得学士学位。职位名称方面，受访者包括一名博士后研究员、一名研究助理、一名研究科学家，其余七人均为博士研究人员。五人表示每天会多次进行学术网络搜索，四人表示每周至少会进行多次，另有一人表示每周进行的频率更低。八人有三年或以上开展系统性文献综述的经验，一人表示有两年经验，另一人则有一年经验。

参与者来自多个不同领域：神经科学、地理学、生物医学、人机交互、自然语言处理、增强现实/虚拟现实设计以及可穿戴计算。他们表示，探索文献时主要使用学术搜索引擎和论文列表，包括 Google Scholar、Semantic Scholar，以及特定领域的会议论文集、期刊和组织（例如，自然语言处理领域的 ACL 或人机交互领域的 CHI'23 会议论文集）。他们使用各种应用程序来阅读和撰写科学文献。

我们询问参与者进行文献综述时的工作流程，以及他们遇到的挑战。随后，我们给他们 20 分钟时间，探索并阅读从多篇关于某一主题的研究论文中摘取的一组相关工作段落。这些段落以列表形式显示在一个简单的文本搜索引擎界面中（见附录图 8）。为了让探索过程更有背景，我们给他们安排了一项模拟任务 [4]：开展初步研究，以大致了解"错误信息、假新闻和事实核查"这一主题，最终目标是撰写文献综述。为了深入了解他们的用户体验，我们请参与者在探索段落列表时出声思考。之后，我们询问他们阅读多篇段落而非论文的体验、其中面临的挑战，以及他们用来克服这些挑战的策略。接着，我们向他们展示了其他原型设计：在相关工作段落中高亮显示尚未探索的参考文献和术语，并弱化冗余引文。

访谈由第一作者通过视频通话远程进行，时长约为 45 分钟。访谈经过录音，随后使用自动转录服务转写。接着，第一作者审阅了转录文本，并采用开放式编码方法对其进行编码，归纳主题 [11]。经过多轮迭代，并定期与研究团队其他成员讨论，我们确定了用户挑战，进而制定了我们的方法的设计目标。

3.2 文献综述中的用户体验

3.2.1 当前的文献综述工作流程与挑战。在询问他们目前的工作流程时，所有参与者（10/10）都提到会使用学术搜索引擎来查找某个主题的相关论文。一些参与者（5/10）提到会通过协作者、导师或 Twitter 等社交媒体共同收集论文清单。他们都提到会逐篇阅读论文以提取含义；3/10 提到会在 PDF 文档上添加笔记和高亮标注；2/10 提到会将这些论文保存到书目管理器（如 Zotero、Mendeley）中。

谈到挑战时，10/10 的人都提到，很难在不同论文之间建立联系。8/10 的参与者提到了面对未知未知时的挑战，范围从不知道哪些搜索关键词能找到合适的论文，到不清楚感兴趣的主题中包含哪些潜在子主题："我常常不知道该搜索哪些关键词/领域专用语言，才能找到合适的文献"，"即使有人推荐给你一份论文候选列表，要全面了解这个主题也很难，而且我觉得自己可能会目光短浅，存在盲点"。

7/10 的人提到，很难记住自己之前读过哪些内容：“很难跟踪许多不同的研究脉络和观点，也难以把握全局。” 5/10 的人谈到了难以确定阅读优先顺序：“当我看到结果中有这么多论文时，会不知所措，于是把它们都打开到不同的标签页里。但之后我不知道该先读哪篇，所以这些标签页就一直开着。”。

3.2.2 探索相关工作章节的偏好与挑战。当被问及阅读多个段落以获得主题概览的体验时，10 位参与者中有 9 位更喜欢阅读相关工作段落，而非论文。一些积极反馈提到了由此获得的主题整体概览：“即使只读几个段落，也能让我迅速掌握该主题的相关词汇、参考文献和要点”，“我能够立即看出不同的研究脉络”。其他人谈到了文本对被引用论文进行总结的价值：“我喜欢参考文献周围的补充说明，这样我就能理解背景，并迅速决定是否要打开论文继续阅读”。还有一些人间接提到，他们已经会借助论文来寻找其他值得阅读的论文，而提取相关工作章节及其中的引文能让这一过程更加顺畅：“这肯定比阅读 *PDF* 文件更有帮助，也省去了那套流程：打开 *PDF* 文件、阅读引言和论文正文，再在参考书目中的参考文献与论文之间来回查找，以确定哪些论文可能有用”。

不过，阅读多个段落以获得主题概览也面临一些挑战。一些参与者希望获得优先级排序和导航支持，以便知道接下来该读什么：“难以确定这些内容的阅读顺序” (5/10),“不清楚段落之间有何异同” (7/10), “想知道本段总结了哪些最重要或最核心的论文” (3/10). 参与者还希望获得支持，以跟踪自己的探索进度：“ 想记录哪些 *[* 论文和段落 *]* 已经读过，哪些还没读 ” (2/10), “难以判断这个主题还有多少内容需要阅读” (2/10).

当被问及希望如何确定段落和论文的阅读优先级时，参与者提到，他们希望优先阅读覆盖面广的段落；这些段落摘自涵盖近期研究和基础性研究的论文，且这些论文应为高被引论文，最好是综述论文，同时尽量少讨论论文自身的研究工作。此外，在排序方面，大多数参与者（8/10）表示，最先阅读的几个段落应能展现子主题的多样性；另有 6/10 的参与者表示，相似段落不应出现在同一页面上。

3.3 设计目标

基于上述研究发现，我们列出了核心设计目标：

[D1] 帮助用户推断主题更高层次且有意义的组织结构，并深入探索子主题**[D2]** 使用户能够灵活地确定优先级，并探索相关工作段落之间的相似之处和差异**[D3]** 帮助用户跟踪已探索的段落和参考文献

4 Relatedly 系统

在形成性研究所得见解的指导下，我们开发了 Relatedly，这是一种新颖的文献综述方法，可帮助用户更全面、更深入地了解某一研究主题。

本节将通过一个用户场景示例、对系统主要功能的演示、对自动章节标题生成流程的说明与评估，以及对整个系统实现细节的介绍，来描述该系统。

4.1 用户场景示例

设想一位初级计算机科学研究人员，她希望广泛了解一个自己并不熟悉的研究领域——错误信息与事实核查。由于不知道有哪些重要的子主题，她首先使用常见的学术搜索引擎开展文献综述，并搜索“错误信息”和“事实核查”这两个短语。然而，尽管搜索结果中的所有论文看起来都相关，但仅凭论文标题和搜索摘要，她很难看出更高层次的主题，以及单篇论文之间的关联。

她感到不知所措，于是使用相同的查询切换到 Relatedly，系统随即在概览视图 中返回与查询相关的相关工作段落列表（图 1）。为了了解整体情况，她快速浏览章节标题，很快便了解到不同的子主题，例如事实核查数据集、社交媒体、新闻与错误信息和假新闻检测技术。章节标题让她能够快速浏览概览视图，了解不同的高层次研究重点。作者通常会根据主题将相关工作部分划分为相关的小节，因此这些标题有助于概括段落的关注重点。

她开始对社交媒体、新闻与错误信息这一子主题感兴趣，于是开始阅读以此为章节标题的相关工作段落。她点击一些参考文献，查看其元数据，包括标题、摘要、TLDR、 [8], 作者、发表年份、会议和引用次数，并点击“复制”来收集一些想稍后阅读的论文。注意到当前段落发表于四年前后，她点击“探索相似段落”按钮。在相似段落视图 中（图 2），系统会显示搜索结果中其他同样涉及社交媒体、新闻与错误信息的段落。浏览这些相似段落时，她了解到其他作者如何在不同段落中概述这一特定子主题的先前工作（例如《网络新闻业的模糊边界》、《事实核查的两难》和《事实核查声明》），这些段落摘自不同的来源论文。她开始理解多篇被引用论文与该子主题中讨论的概念之间的联系。她逐渐觉得自己对这一子主题有了更全面、更完整的认识。她继续阅读其他段落，包括近期发表的段落，以便全面探索这一子主题。

随后，她返回概览视图，探索新的子主题。她注意到，经过迄今为止的探索，展示给她的段落已经发生了变化。段落会动态重新排序，优先显示包含更多尚未探索且差异较大的参考文献的段落。她还注意到，有些段落中，提及她已经探索过的论文的句子被淡化显示。现在，有时段落中的某些行内引文也会高亮显示，这些引文指向尚未探索、且在语义上与她之前探索过的参考文献不同的文献。她

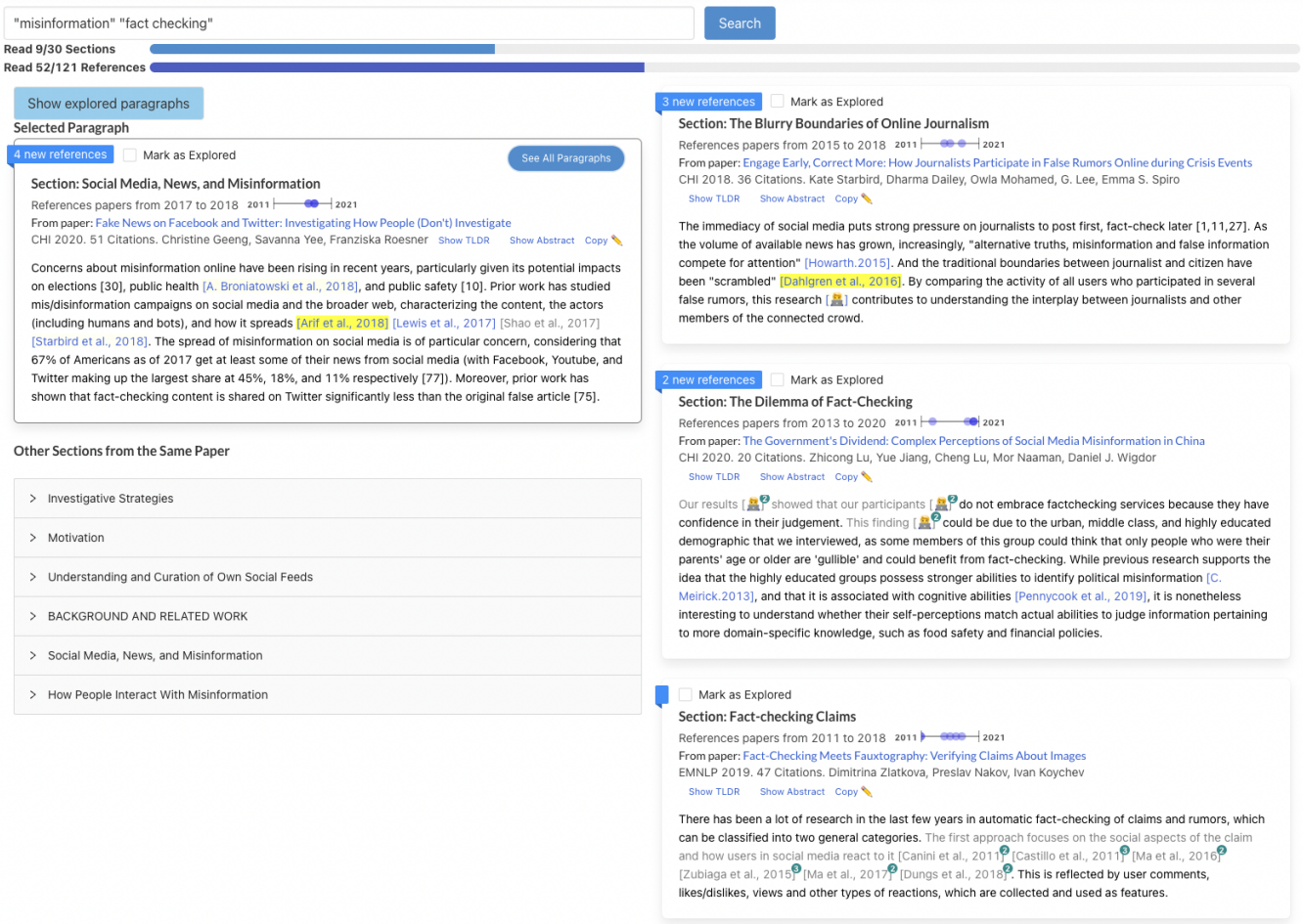

图2：若要进一步阅读概览视图（图1）中特定段落所讨论的子主题，用户可以通过相似段落视图探索该子主题中引用了相同或相似参考文献的其他段落。

略过了一些有许多句子被淡化显示的段落，转而关注一个有多处高亮的段落。最后，她查看进度条，以跟踪她迄今为止探索了全部相关工作段落和参考文献的多少。

4.2 系统功能

我们根据形成性研究提出的三个主要设计目标来组织系统功能的描述。

4.2.1 [D1] 推断主题概览 + 深入探索子主题。 给定搜索查询后，Relatedly 会在概览视图 中呈现与查询相关的相关工作段落列表（图 1）。每个段落都代表一个不同的子主题，并经过排序，以涵盖广泛的子主题。若要深入探索某个段落所涵盖的子主题（例如《事实核查声明》），用户可以点击该段落卡片右上角的"探索相似段落"按钮。这会打开相似段落视图，在右侧栏中显示相似段落，并将选中的段落固定在左侧栏（图 2）。以下介绍用于生成主题概览并深入探索子主题的具体功能。

段落多样性排序：为了帮助用户在概览视图中了解高层次的组织结构，Relatedly 首先根据标准 BM25 [60] scor- ing,[4] 检索最相关的段落，然后使用最大边际相关性（MMR）技术 [9] 对检索到的段落重新排序，以平衡排名靠前的结果与查询的相关性和信息新颖性。原始 MMR 技术依靠文本相似度来衡量给定文档的信息新颖性，而这里我们使用每个段落中尚未探索的参考文献数量来近似其信息新颖性。目标是重新排列段落，使排名靠前的段落共同包含数量最多的独特且尚未探索的参考文献，并呈现多种多样的子主题，同时兼顾它们与查询词的相关性以及参考文献数量。此排序基于

[4]为控制用户研究的时长，我们默认展示排名前30的段落。

参与者在形成性用户研究中对段落排序方式的反馈。排名为 i 的段落的排序分数如下：

$$MMR_i = \arg\max_i \{BM25_i[\lambda|\text{参考文献}_i| - (1-\lambda)|((\cup_{j=1}^{i-1} \text{参考文献}_j) \cup \text{参考文献}_{exp}) \cap \text{参考文献}_i|]\}$$

在该方程中，$BM25_i$ 是基于段落文本和查询词计算的相关性分数，Refs_i 是按位置 i 排序的段落中的参考文献集合，Refs^{exp} 是用户已经浏览过的参考文献集合，而 λ 是一个超参数，用于调整因包含已出现在先前排序段落中的参考文献而产生的惩罚项。[5]

描述性段落标题：概览视图中的每个段落都配有一个描述性标题，用于概括该段落的重点，并作为用户的子主题。由于作者通常会根据主题将相关工作部分组织成相关的子章节，因此对于大多数段落，这些标题都取自相关工作章节的章节标题。对于作者仅撰写了信息量较少的通用或简短章节标题（例如"相关工作"或"事实核查"的段落），或没有章节标题的段落，Relatedly 会使用基于 BART 的 [42] 模型生成描述性标题（详见第 4.3 节）。

给定一个段落时的相似段落：在相似段落视图中，会为选定段落显示一份相似段落列表。我们根据这些段落是否引用了与选定段落相同的论文（主要排序依据），或者是否在语义上相似（次要排序依据；通过 SPECTER 论文嵌入之间的欧氏距离阈值来判断 [13]），来确定段落相似度。

阅读给定段落所属的论文：在相似段落视图中，用户可以在选定段落下方查看同一篇论文中的所有其他部分，包括相关工作部分的其他内容，以便更全面地了解该段落的背景。

4.2.2 [D2] 优先处理并探索相关工作段落中的异同。在形成性研究中，参与者表示，"很难决定按什么顺序阅读这些段落"。因此，Relatedly 提供了多项功能，以支持用户对多样化的未探索信息进行优先排序和阅读。

未探索参考文献计数徽章：形成性访谈和先前工作都表明，用户希望优先阅读未读信息最多的段落。为此，所有段落都带有未探索参考文献计数徽章（例如 11 new references）），用于显示本段落讨论的尚未探索的唯一论文数量。随着用户与各段落中的更多参考文献互动，该数字会动态更新。排序算法会优先将包含更多未读参考文献的段落排在概览视图的较高位置。

高亮显示尚未探索且彼此不同的参考文献：为了进一步帮助用户优先关注尚未探索的新信息，并满足"找出论文之间的异同"这一需求，Relatedly 会高亮显示尚未探索且彼此不同的参考文献（例如 [Harrington et al., 2021]）。随着用户点击并阅读参考文献和段落，部分参考文献会以黄色高亮显示，这表示这些论文在语义上与用户此前交互过的其他论文不同（根据其 SPECTER 论文嵌入之间的欧氏距离是否超过阈值来计算 [13]）。这些参考文献会以渐变的黄色高亮显示：黄色越亮，被高亮的论文就与用户此前交互过的最相似论文差异越大。

参考文献时间线可视化：用户还希望采用另一种启发式方法，优先查看同时涵盖该主题近期研究和基础性先前研究的段落。为帮助用户进行初步筛选，所有段落都配有参考文献时间线可视化，用于展示本段落所引用论文的时间范围（如 2011 2021）。其中，一篇被引用论文的出版年份。由于时间线的最小值和最大值分别表示所有段落所引用论文中最早和最晚的年份，用户可据此优先阅读引用近期论文或较早基础性论文的段落。这项功能源自形成性研究中参与者的反馈：他们希望根据段落所引用论文的新近程度来确定阅读优先级。

引用频次徽章：形成性访谈中的参与者提到，他们希望优先关注段落中的某些部分，尤其是当某个论点或段落引用了多篇论文时，想知道应该优先阅读哪篇论文。Relatedly 提供引用频次徽章，旨在表明某篇参考论文对某个主题的核心程度或重要性。这些带数字的绿色标签表示这篇论文在这些结果段落中被引用的次数（例如 [Popat et al., 2017]）。如果概览视图返回的多个段落都引用了这篇论文，就意味着多位作者讨论了这篇论文，因此它可能是该主题的核心或重要内容。

自我引用图标：形成性访谈中的参与者提到，他们希望降低段落中某些部分的优先级，尤其是作者在背景部分介绍自身研究的内容。为此，Relatedly 会识别段落中哪些部分提到了论文的研究，并通过图标向读者标示。

4.2.3 [D3] 跟踪对论文和段落的探索进度。除了希望优先关注尚未探索且内容不同的信息外，用户还提到，难以跟踪哪些论文或段落已经读过、哪些还没有。Relatedly 提供了多项功能，让用户了解自己对内容的阅读进度，同时尽量减少重复。

弱化显示先前遇到的信息 Relatedly 通过将段落中的整个句子变灰，弱化显示先前遇到的信息。如果用户点击了段落中引用的一篇论文，而该论文也在另一个尚未阅读的段落中被引用，那么该处的参考文献及对应文本也会被弱化显示，以表明用户先前已经遇到过这些信息。

将段落标记为已探索：同样，在段落级别，一旦某个段落被"标记为已探索"，它就会从概览视图中移除。用户可以点击页面左上方的"显示已探索段落"按钮，查看这些已探索的段落。此外，每当一个段落被标记为已探索时，

[5]根据开发过程中的试玩测试，我们将 λ 设为 0.3，以适度优先包含更多参考文献的段落，并对包含已被排名较高段落涉及的参考文献的段落施加较高惩罚项，从而使排名靠前的结果覆盖更多不同主题。

其中的所有参考文献都会被视为“已遇到”，并且段落列表会根据其中未探索论文的数量、该段落与已阅读内容的差异程度，以及与所查询主题的相关性进行动态重新排序（公式 1，详见前文）。

进度条：随着段落被标记为已探索，页面顶部的段落进度条也会更新（如……）。随着用户点击参考文献，页面顶部的论文进度条也会随之更新，并显示用户已阅读了 n 篇论文，占返回段落中唯一参考文献总数的比例（如……）。各段落中的未探索参考文献计数徽章也会更新。此状态会在不同查询之间持续保留，因此用户发起新查询时，如果之前读过某些论文或段落，进度条也会记录这些阅读进度。此功能旨在应对用户面临的一项挑战：难以追踪已阅读论文中的信息以及已探索的子主题。

4.3 自动章节标题生成

Relatedly 利用相关工作章节中的章节标题，为用户快速概览不同的研究脉络。其中一项挑战是，并非所有作者都会创建描述性小节标题。在这种情况下，对于假新闻 查询，显示通用的上级章节标题（例如“相关工作”或“假新闻”）对用户帮助不大。为此，我们开发了一种自动标题生成模型，该模型使用启发式方法识别出的描述性章节标题进行训练，并将其应用于来源论文中作者未撰写描述性标题的段落。

4.3.1 方法。 我们尝试了两种常用的基于 Transformer 的序列到序列模型来生成标题：(i) BART [42], 和 (ii) T5 [59]。由于性能出色且能够适应不同任务，这些预训练模型已成为开发各种文本生成模型的事实标准起点。为了进一步训练这些模型，使其能够生成科学文献标题，我们采用一组启发式规则，从数据集中收集包含描述性标题的段落。这些规则包括筛除以下标题：仅由首字母缩略词构成、少于三个词，或包含通用术语。[6] 这种严格的筛选倾向于优先保证精确率而非召回率，以减少筛选后的数据集中用于训练和测试的噪声。最终数据集包含 23,957 个段落及其标题，我们又将其随机划分为 80% 的训练集、10% 的验证集和 10% 的测试集。我们在训练集上对 BART 和 T5 的大型版本进行 10 个轮次的训练，并根据验证集上的损失选择绩效最佳的模型检查点。表 1 使用 ROUGE指标 [43], 对两个模型在留出测试集上的表现进行了评估；该指标通过计算生成标题与人工撰写的标题之间的 n元语法重叠来衡量标题质量。根据这些分数，两个模型生成的标题质量似乎相近。我们抽取并检查了一小部分生成的标题，对两个模型进行了比较，发现 BART 模型生成的标题往往更详细且更具描述性。因此，随后我们使用 BART 模型为筛选集之外、没有描述性作者生成的标题的段落生成标题（共 39,186 个，约占整个数据集的 62%），以便在下一小节中进行更严格的人工评估。

模型	ROUGE-1	ROUGE-2	ROUGE-L
BART	30.7	14.2	28.8
T5	31.0	14.3	29.0

表 1：两个模型在描述性章节标题测试集划分上的 ROUGE 分数。ROUGE-1、-2 和 -L 分别衡量生成标题与作者撰写的标题之间的一元语法、二元语法和最长子序列的重合度。

4.3.2 人工评估。我们对基于 BART 的标题生成模型进行了人工评估。这一点很重要，因为 ROUGE 等自动评估指标并不总是与人类感知高度相关。我们手动评定了四组标题：针对训练中未包含描述性标题的段落（测试集划分）和不含描述性标题的段落，分别评定作者撰写的标题和模型生成的标题。我们采用五点同意度对以下两项陈述进行评分：标题具有描述性，并具体体现了某一研究脉络（S1）和标题涵盖了段落中的所有要点（S2）。S1旨在衡量标题的质量和具体程度，S2则旨在衡量标题对段落文本的呈现程度。

为确保评分质量，两位作者进行了两轮重复评分，每轮随机抽取40个标题（每组10个）。经过两轮比较和讨论，以校准评分标准后，基于 Krippendorff α系数计算的标注者间一致性在 S1 和 S2 上分别达到 0.90 和 0.74。[7] 随后，作者对400个标题（每组100个，成对呈现）进行了非重复评分。评分过程中，作者并不知道每个标题取自哪种条件，以避免偏差。

如图3所示，我们的模型生成的标题质量可与作者撰写的较长且描述性强的标题相媲美（S1：p=0.40；S2：p=0.32；n = [100, 100]，Wilcoxon 符号秩检验）；而在作者未提供描述性标题时，模型生成的标题质量则显著更高（S1：$p<0.001^{***}$；S2：$p<0.001^{***}$；n = [100, 100]，Wilcoxon 符号秩检验）。除了这些评分，我们还快速检查了所有模型生成的标题（共200个），以筛查重复和幻觉，即提及段落中未出现的概念。我们这样做是为了进行合理性检查，因为生成模型往往容易出现这些问题，尤其是幻觉 [31]。根据筛查结果，我们发现模型很少出现这些问题——仅有5/200个（2.5%）标题存在重复，9/200个标题（4.5%）存在幻觉。这可能部分是因为生成的标题较短，因此重复和幻觉出现的空间较小。鉴于这些令人鼓舞的结果，在系统中，如果某个段落没有作者撰写的描述性标题，我们就会显示模型生成的标题（图1）。以下是一些

[6]文献综述、背景、局限性、未来工作、结论、讨论、相关工作、结果。

[7]第1轮的协议度分别为 S1 的 0.66 和 S2 的 0.49。

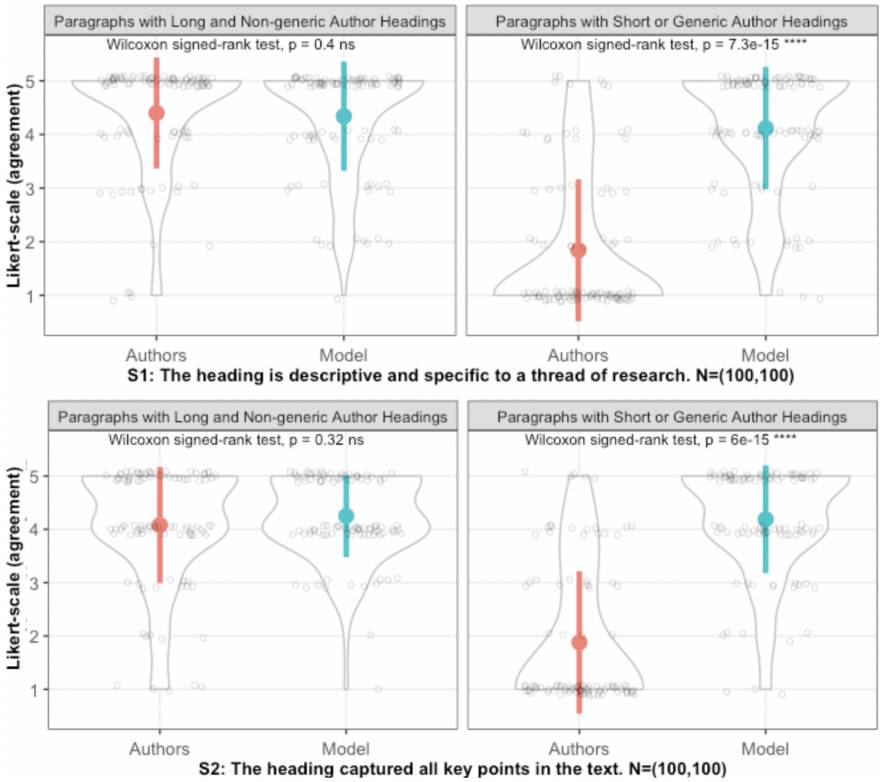

图 3：对模型生成和作者撰写的章节标题进行人工评估。结果表明，当作者撰写的标题较长且不笼统时，模型生成的标题质量与之相当；而当作者没有撰写此类标题时，模型生成标题的质量则显著更高。

模型生成的标题及其对应的作者撰写的标题并列示例见附录（表3和表4）。

4.4 实现细节

Relatedly 是一个标准的 Web 应用程序。前端使用 ReactJS 框架，以 TypeScript 编写，约有 3,500 行代码。后端使用 Python 和 SQL 编写，约有 1,500 行代码。我们使用 Flask 作为 HTTP 服务器框架，并使用 PostgreSQL 数据库来访问数据集，以及记录用户研究中的行为日志。我们使用 Whoosh[8] Python 库实现标准的 BM25 文档检索算法 [60]，，以支持对段落进行全文搜索。在评估研究中，系统的所有交互（例如新查询、阅读的论文和段落等）都会以 JSON 格式记录到数据库中（参见补充材料）。服务器、数据库与用户浏览器之间的所有通信均经过加密，并通过创建匿名会话和用户 ID 进行匿名化处理。

4.4.1 数据集。为了测试 Relatedly 方法，我们从 S2ORC 中收集了一个完整论文数据集，涵盖 2016—2021年 期间发表于五个 人机交互 和 自然语言处理 会议（ACL、EMNLP、UIST、CSCW、CHI）的论文。S2ORC 是一个大型开源语料库，包含 8,110 万篇涵盖众多学科的英文论文 [44]。这些主题的选取主要是为了方便作者在开发过程中评估该系统的实用性，也为了在用户研究中招募可能对这一主题更感兴趣的参与者。为了找到总结多项既有研究的段落，我们针对每篇论文提取了所有包含三条或更多参考文献的段落及其章节标题。由于相关工作部分通常会引用其来源论文，而单独阅读时这类引用可能显得缺乏上下文，因此我们使用一个简单的词表，将自我指涉短语（例如，本文、我们的方法、我们的系统，等等）解析为来源论文。最终，该数据集包含从 11,382 篇论文中提取的 63,144 个段落。其中约 49,975 个段落来自相关工作章节，其余段落主要提取自引言和讨论部分。S2ORC [44] 将文内引用解析为其元数据，包括作者、引用次数、摘要和简明摘要 [7]。这也使 *Relatedly* 能够在系统中将参考文献格式化为 APA 格式（即第一作者的姓氏和出版年份），从而让相同参考文献在不同段落中呈现为一致的形式。

[8]https://github.com/mchaput/whoosh

5 用户评估研究设计 N

Relatedly 的设计改变了常见的文献综述流程：不再探索单篇论文（例如通过搜索引擎查找），而是探索介绍某一主题下多篇论文的段落。为了

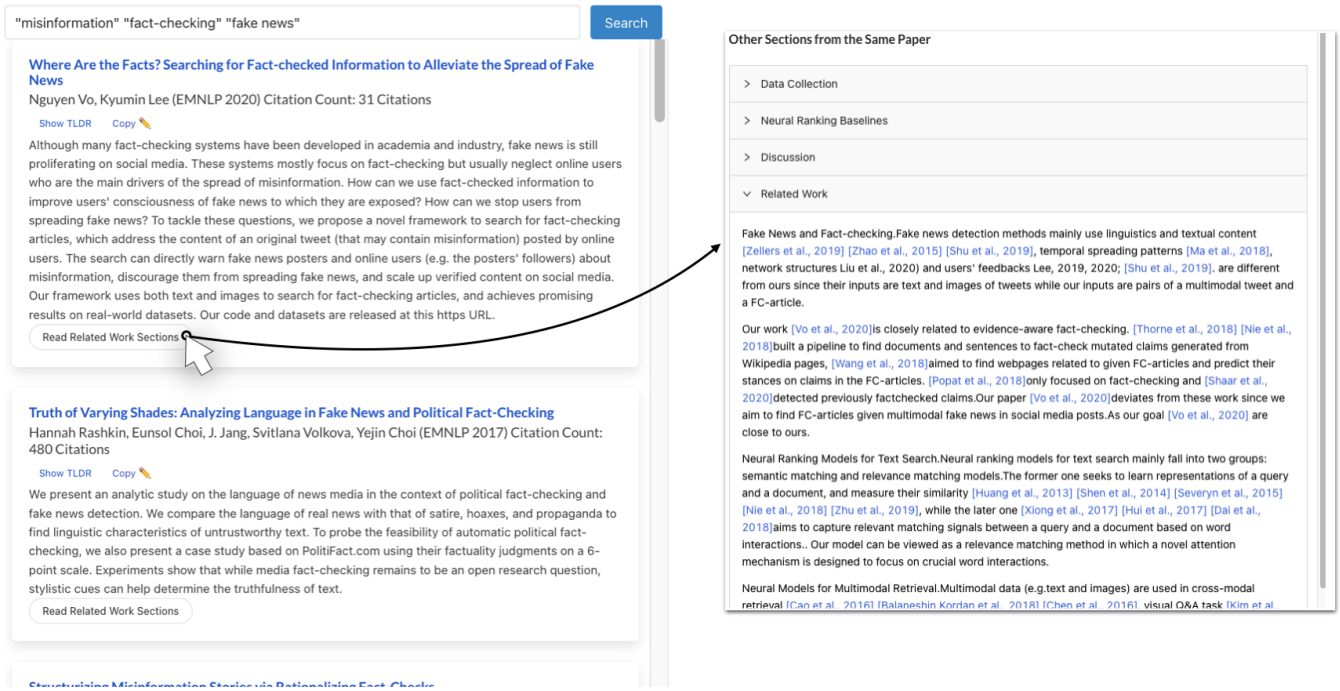

图 4：模拟常见学术搜索引擎的基线条件（左）。此外，“阅读相关工作章节”按钮允许用户以较低的交互成本阅读论文中包含三个或更多参考文献的段落（右）。

探究其影响，我们开展了一项被试内实验，邀请 15 位参与者进行文献综述，并将*Relatedly* 与作为基线的标准论文搜索引擎进行比较。研究期间，参与者使用分配给自己的系统探索文献，目标是为撰写一篇关于指定主题的综述论文创建提纲和笔记。这使我们能够了解参与者在任务过程中学到了什么。研究结束后，我们分析了行为日志，以了解他们如何使用各个系统，并评估提纲的质量，从而判断哪种流程能帮助参与者更全面地了解文献。

5.1 实验设置

我们将*Relatedly* 与一个模拟标准学术搜索引擎的基线系统进行比较，并采用被试内实验设计。在研究期间，每位参与者都使用这两个系统，针对两个不同主题进行文献综述。为控制个体差异和学习行为的影响，我们对主题和实验条件进行了平衡，以减少顺序效应。

5.1.1 基线条件。基线条件是一个标准的 BM25 搜索引擎，它会返回数据集中标题或摘要中提及查询词的论文列表（图 4）。对于搜索结果中的每篇论文，用户都可以访问其元数据，包括标题、作者、发表场 venue、出版年份、摘要以及 TLDR 总结[8]。[9] 为降低使用基线条件的交互成本，用户可以点击“阅读相关工作章节”按钮，以访问包含三个或更多参考文献的章节标题和段落。这样可以确保：1）参与者在两种条件下都能访问相同的数据；2）在基线条件下，访问这些数据的交互成本较低。与 *Relatedly* 条件类似，参与者还可以点击“复制”按钮，复制论文标题并将其粘贴到提纲中。

5.1.2 任务。 为了让参与者的探索和意义建构置于具体情境中，我们为他们提供了一个模拟工作任务情境 [4]，要求他们开展初步研究，了解主题的大致情况，最终目标是撰写一篇综述文章：

> 设想你正在调查并总结人机交互和自然语言处理领域关于以下主题的科学研究：[两个任务主题之一：错误信息、假新闻与事实核查，或众包]今天，请先开展初步研究，以全面了解该主题。你的目标是全面了解这一主题，并通过在此搜索引擎上搜索和收集信息，尽可能多地找出与主题相关的术语、概念和观点。在任务过程中，请在提供给你的笔记文档中撰写一份提纲，以便将来继续完成这项任务。提纲可以包括规划论文的所有章节、记录你找到的重要论文和研究等。

作为评估两种条件下用户行为的被试内研究设计的一部分，每位参与者都完成了

[9]这些是 1–2 句的摘要，在常见的学术搜索引擎上也可以找到。

上述任务两次（即每种条件各一次）。为防止学习产生的迁移效应，每位参与者都针对以下列出的两个主题完成了任务。为避免顺序效应，系统采用了平衡安排，使参与者在每种条件下看到的主题各不相同。

- **错误信息、假新闻与事实核查：** 互联网让数十亿人只需敲击几下键盘，就能轻松获取信息。然而，互联网也让虚假信息更容易传播，从而对个人乃至整个社会造成灾难性影响。人机交互和自然语言处理领域的研究聚焦于检测方法及其应用和影响。研究假新闻的影响，以及为应对假新闻而开发的方法。
- **众包：** 众包是指大量分散的参与者共同完成一项任务。随着我们迈向由数字平台推动众包的新工作形态，人机交互和自然语言处理领域的研究聚焦于各种众包方法及其在不同领域中的应用。研究众包的方法及其应用。

所选任务主题涵盖的信息空间相对广阔、复杂且多面，普通人在开始任务时对相关内容的了解通常较为有限。这些任务也具有相当强的跨学科性质，因此即使参与者具备相关领域的专业知识，在这一领域仍有更多可学之处。此外，这些主题在我们的数据集中也有充分体现——该数据集由人机交互和自然语言处理领域会议的论文构成。

5.1.3 参与者。 我们从三所大学的研究实验室招募了 15 名参与者（8 人自我认同为女性，其余为男性；年龄：19–32 岁，M=25.88，SD=3.72）。所有研究均通过视频通话远程进行。参与者完成这项 90 分钟的研究后可获得 45 美元报酬。参与者主要是从事研究工作的研究科学家、博士后和研究生。

5.2 研究流程

研究预约开始前，我们向参与者发送了知情同意书，并请他们填写人口统计信息。在研究预约期间，每位参与者都完成了两项文献综述任务；任务顺序以及任务与系统的分配组合均经过平衡。每项任务持续 20 分钟。在这 20 分钟内，参与者可以自由地与系统交互，并在 Google 文档中创建学习提纲，同时大声讲述自己的体验 [28, 41]。开始每项任务前，参与者会观看相应系统的简短教程，并获得 5 分钟时间，使用“意义建构”这一测试主题探索系统。

5.3 测量指标

5.3.1 学习提纲的质量。 我们的主要测量指标聚焦于通过分析参与者在实验室研究中撰写的学习提纲，评估*Relatedly* 在支持文献综述方面相较于基线的表现。为此，我们请主题专家在不知道各提纲所属实验条件的情况下，检查并评估每份 Google 文档提纲。我们将专家定义为：专注于该主题并获得博士学位，且在该领域拥有多年经验并发表过多篇论文的人。在研究所用任务中，两位作者符合上述标准，且不知道实验条件。专家统计了参与者在提纲中添加的研究主题数量（作为全面性的替代指标），并从以下方面对提纲进行五分制评分（分数越高越好）：

- 连贯性：类别结构是否合理，以及其中的论文和子类别是否契合。
- 洞察力：类别是否具有洞察力，并捕捉了该领域中的重要研究脉络。
- 细节程度：类别是否包含相关子主题和论文的丰富细节。

这些标准的制定参考了关于聚类人工评估的文献（例如 [73]）以及自然语言处理领域针对自动生成提纲的评估标准（例如 [21]）。两位专家进行了两轮评分和讨论，以校准主题数量及其最终得分。随后，将这些得分相加，用于计算提纲的总体质量。

5.3.2 行为日志分析. 我们利用两种条件下的行为日志，衡量每位参与者在段落和论文层面与两个系统交互的频率。例如，点击参考文献、复制论文标题或浏览段落。此外，我们还考察了参与者使用仅在*Relatedly* 条件下提供的功能的频率，例如参考文献高亮和淡化、引用频次徽章，以及概览视图和相似段落视图。

5.3.3 定性洞察与感知价值。 为了更深入地了解使用这两个系统所面临的挑战与带来的益处，我们转录了参与者在完成任务期间的出声思考记录。随后，第一作者采用开放式编码方法 [11]，分两轮阅读了转录稿。通过与研究团队其他成员讨论，我们确定了参与者经历中的共同主题。此外，我们还开展了任务后调查，请参与者使用5点李克特量表，对围绕系统价值的一组陈述进行同意程度评分（表3）。

6 研究发现

本节结合上一节所述的三项测量结果，全面呈现使用 *Relatedly* 相较于基线条件的成本与收益，其中基线条件模拟了标准学术搜索引擎。

6.1 更高质量的综合提纲

根据参与者对三个方面的 5 分制专家评分总和，与基线相比，使用 Relatedly 时写出的提纲质量更高（M = 10.53, SD = 3.64 对比 M = 8.07 SD = 4.61 t = 2.58 p = 0.02 15, 5 , ; , 图中）。为校正多重比较，我们进行了 Wald 型检验和重复测量多元方差分析，发现两个系统在连贯性、洞察力和详细程度的综合测量上存在显著差异（$manovaF$ = 7.78, p = 0.02*, $Wald\chi^2$ = 13.18, p =

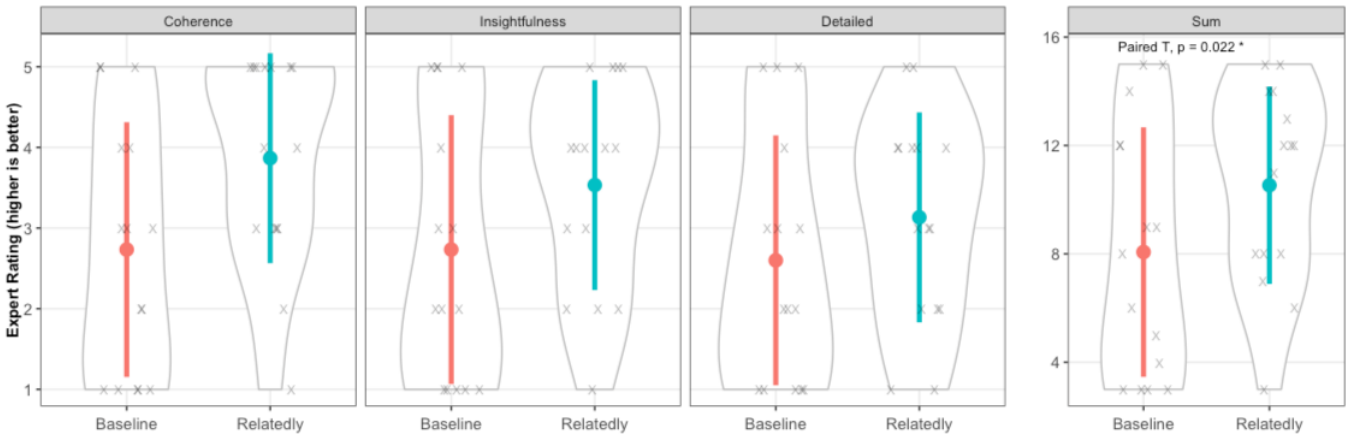

图 5：参与者分别使用这两个系统进行 20 分钟的文献综述后，生成了学习大纲摘要。专家根据 3 项标准，使用 5 点李克特量表对摘要进行评分，以表示认同程度（5 表示非常认同）。评分时，专家不知道每份大纲来自哪种条件。研究采用多元方差分析和 Wald 型检验校正多重比较，发现两种条件在连贯性、详尽性和洞见性这三个综合因变量上存在统计学显著差异（$manovaF = 7.78$，$p = 0.02^{*}$，$Wald\chi^{2} = 13.18$，$p = 0.004^{**}$）。

0.004**；图 5）。[10] 进一步细分这项质量得分后发现，与使用基线相比，参与者使用 Relatedly 时写出的提纲连贯性显著更高（$M = 3.87$，$SD = 1.30$，满分 5 分），洞察力也显著更强（$M = 3.53$，$SD = 1.30$，满分 5 分）（连贯性: 2.73 $SD = 1.58$ 5 $M = 2.73$ $SD = 1.67$ ，满分为；洞察力：，满分 5 分）。此外，参与者使用 Relatedly 时写出的提纲也更加详细（$M = 3.13$，$SD = 1.30$ 满分 5 分），而使用基线时则较不详细（$M = 2.60$，$SD = 1.55$ 满分 5 分）。

对出声思维录音的定性分析揭示了参与者使用 *Relatedly* 时所采用的探索策略。最常见的是，15 名参与者中有 13 名谈到采用"广度优先的方法来探索不同的段落和主题"，其中 8 名特别提到使用未探索参考文献计数

段落级 11 new references 用于优先排序。例如，P09 使用概览视图快速了解其大纲中的不同主题和相关论文，并利用*Relatedly*动态排列段落的方式，最大限度地提高边际新颖性 [9]:

> "我大部分时间都在查看所有段落视图中的各种摘要。——我喜欢它们按照未读参考文献数量从多到少排序，所以按顺序查看很合理。我发现新主题时会把它们记下来——有时写成简短的摘要，有时会把整体主旨的完整引文也记下来，然后复制相关论文的信息，以便需要时能回去查阅。"(P09).

所有参与者都会在概览视图和相似段落视图之间切换，以了解整体概况并深入探索不同的子主题。此外，使用 Relatedly 时，参与者在概览视图中探索的段落明显多于在相似段落视图中探索的段落（ $M =$ 5.80 $SD = 8.00$ $M = 1.80$ $SD =$，，而在相似段落视图中为，

[10]R 包：*MANOVA.RM*：基于重抽样的多元数据和重复测量设计分析

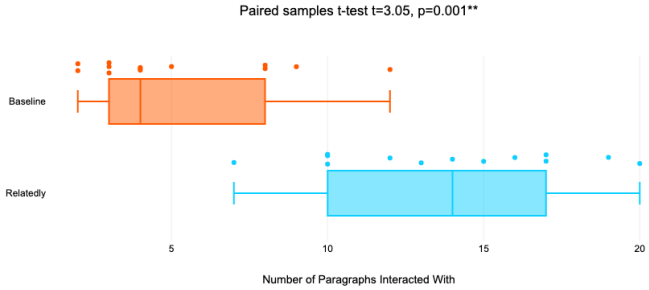

图 6：使用 Relatedly 的参与者比使用基线系统的参与者与显著更多的段落进行了交互

1 = .13，$t = 2.55$，$p = 0.02^{*}$）。两位参与者特别提到，他们会在两种视图之间切换，以同时掌控探索的深度和广度：

> "我会把我的搜索方式描述为一种有限深度的深度优先搜索策略。我从一个较宽泛的想法出发，尝试在一定努力程度内找到与该主题相关的所有内容。相似段落 [视图] 是一个很棒的功能！然后，我会在概览视图 [中切换到另一个子主题]，再重复这一过程。"(P17).

这些系统日志分析和定性洞察表明，参与者使用*Relatedly*时能够更好地掌控自己的探索过程，在探索多个不同子主题和深入探究特定子主题之间灵活切换；这或许解释了他们为何收集到更多子主题，并生成了质量更高的概览式学习大纲。

6.2 论文中心与主题中心探索

虽然两种条件下的参与者都可以查看论文的相关工作段落，但我们发现，使用*Relatedly*时，参与者互动的段落数量是使用

基线时的两倍多（Relatedly：$M = 16.85$，$SD = 7.66$ 对比 $M = 7.21$，$SD = 5.70$，$t = 4.40$，$p = 0.001^{**}$）。参与者描述说，使用基线时，他们的探索策略主要围绕单篇论文展开，而非段落。最常见的情况是，15 人中有 9 人提到，他们主要依靠摘要和 TLDR 简要摘要来决定阅读哪些论文："逐篇浏览，快速扫一遍摘要——如果有趣或相关，就多读一些"（P04）。参与者还提到了其他信号，例如引用次数（15 人中有 3 人）、在摘要中发现陌生术语，以便找到可添加到提纲中的类别（15 人中有 3 人），以及阅读相关工作章节（15 人中有 4 人）。

除了两种条件下探索策略的差异之外，我们还发现，参与者在使用*Relatedly*时积极使用了段落阅读支持功能，这一点得到了行为数据和定性数据的印证。平均而言，点击的引文中有38.73%是此前阅读过、被低亮显示的参考文献；37.84%的点击引文带有高亮标记；10.35%的点击引文带有引文频次徽章。根据定性数据，参与者利用阅读支持提示来决定优先或降低阅读活动的优先级。15名参与者中有12名谈到，他们会依靠黄色高亮参考文献来确定阅读优先级：

> ""哦，我明白了，黄色意味着有些不同，而且黄色越亮，看起来就越吸引我。我想看看与我目前读过的
>
> 论文意见相左的论文" (P04) .

此外，15 位参与者中有 11 位会留意引文频次标签，以寻找该主题中的重要论文：

> "我会寻找引用次数高的文献，因为这类文献往往很受欢迎，也更经典或更具奠基性。" (P03).

关于降低优先级，15 位参与者中有 12 位谈到了他们如何利用低亮度显示的参考文献：

> "当我开始阅读一个段落，看到灰显的文本和参考文献时，就能很方便地知道自己之前在另一个段落中读过这篇论文。我会点击它，确认这是哪篇论文，然后看看这个段落是如何讨论它的。哦，看来这个段落对这篇论文的贡献有些不同的解读，所以我会把这篇论文复制下来。看起来围绕它有一些讨论。" (P09).

最后，参与者提到，在使用 *Relatedly* 时，会通过描述性章节标题探索主题。十位参与者明确提到，章节标题会指出有助于组织知识的主题：

> "这些 *[*标题*]*大致是我会查找的几个宽泛类别，我先浏览标题，并把独特的标题添加到我的笔记中……现在我将开始查找与这个子主题标题类似的具体内容"（P14）。

6.3 参与者更青睐 Relatedly

使用两个系统后，我们询问了参与者的偏好，其中 15 名参与者中的 13 名表示，相较于基线通过阅读论文来开展文献综述的方法，他们更喜欢 *Relatedly* 的方法，即阅读与该主题相关的相关工作段落。尽管偏好基线的两名参与者主要将其描述为熟悉的界面，但偏好 *Relatedly* 的参与者给出了丰富得多的解释：(1) 为原本缺乏结构且复杂的探索性任务提供了良好的结构和组织方式："我可以有一条相对清晰的探索路径。我可以利用其他研究人员总结的内容，因此不需要从零开始。"(P05)；(2) 有助于理解多篇论文之间的联系："对于文献综述来说，了解被引用论文之间的联系非常重要……我发现相关工作章节对此很有帮助，而论文列表就没那么有用了。" (P14)；(3) 为论文提供适量的相关背景信息："提供了更多背景信息，因此需要阅读的论文更少了——需要打开的 *PDF* 文件也更少。" (P13)；以及 (4) 有助于跟踪进度，并确定阅读内容的优先顺序："可以帮助我跟踪自己了解该主题的学习进度。"(P03)。

在一项关于参与者对两个系统看法的后测问卷中，参与者认为，与基线相比，Relatedly 系统在"查找与主题相关的研究""理解术语/概念之间的关系""整合多个来源和不同观点的信息""确定阅读优先级"以及"在文献综述流程中跟踪获取或阅读的信息"方面提供的帮助显著更多（图 7 报告了所有陈述的 Wilcoxon 符号秩检验结果）。

最后，为了衡量使用这两个系统时感知到的工作负荷，我们使用了 NASA-TLX 问卷，发现两者之间没有显著差异（附录表 5）。这一结果表明，尽管*Relatedly* 包含的功能显著更多，参与者却没有感到工作负荷更高，而且他们能够利用这些功能，综合出质量显著更高的学习提纲。

6.4 在真实环境中自愿使用

一项有趣但原本未计划的观察是，五位参与者计划为即将提交的论文撰写文献综述，并表示有兴趣在研究结束后使用 Relat- edly。我们将此视为一个机会，以便更好地了解 *Relatedly* 在较长时间内执行现实任务的表现。因此，我们继续允许他们通过自己的电脑访问 Relatedly，以完成各自的任务，并安排在两周后对他们进行访谈，了解他们的用户体验。随后，第一作者对这些访谈的文字记录进行了开放式编码，以归纳共同主题。所有参与者都是来自三所大型研究型大学的博士生，平均年龄为 25.7 岁。其中两人认同为男性，三人认同为女性。表 2 展示了他们现实任务的概况，以及他们使用 Relatedly 的情况。

这些学者表示，相较于阅读单篇论文，他们更喜欢在 Relatedly 阅读相关工作章节，以此来"了解一个主题"（P09），以及"在撰写论文时，确认自己是否涵盖了所有重要的研究脉络、论文和观点"（P13），或"发现该主题的核心论文"（P15）。他们还提到，使用传统的学术搜索引擎查找特定论文或作者仍然很有帮助，并认为这两种方法相辅相成。P13 将自己的体验与以往的文献综述流程进行比较，总结道：

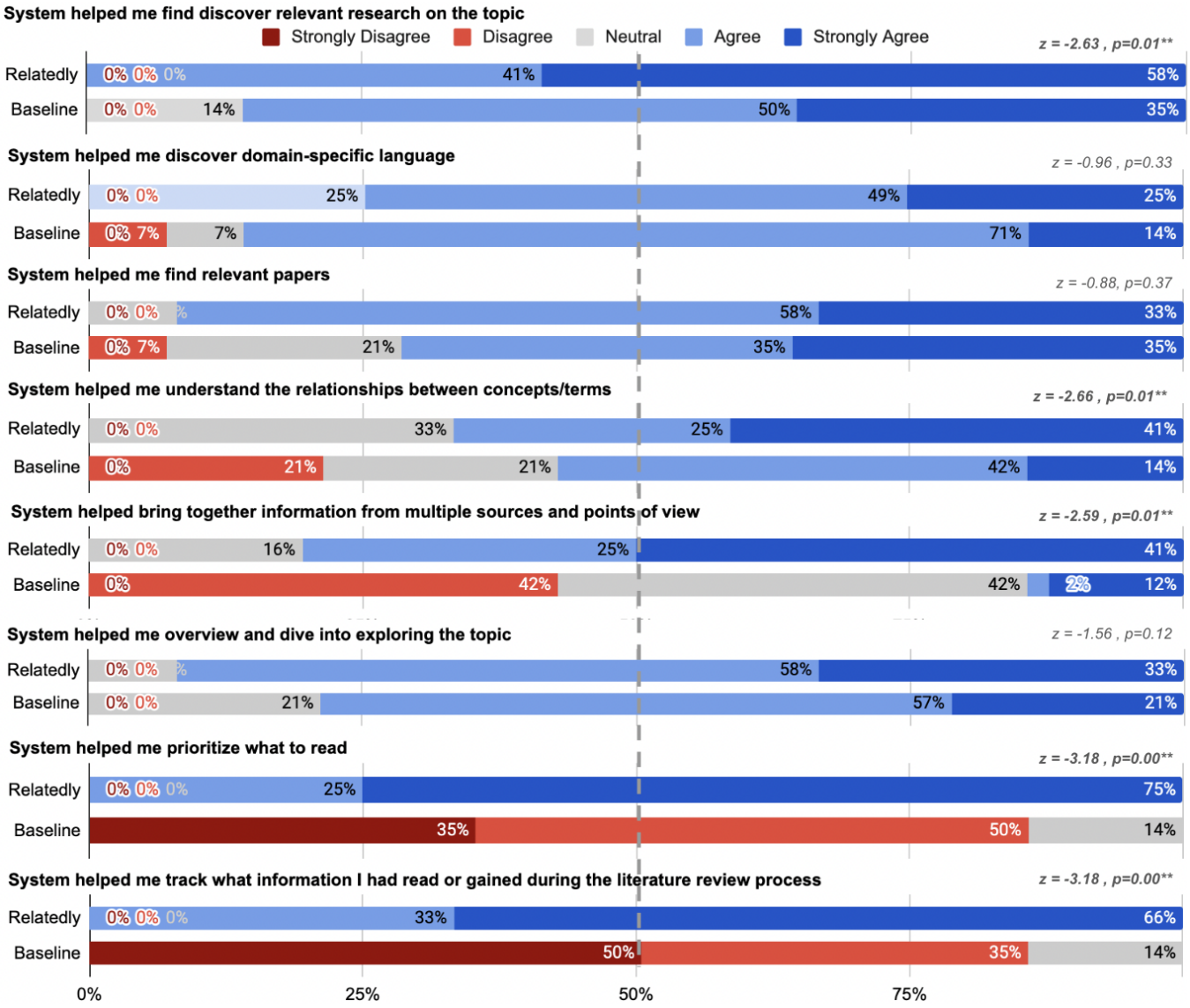

图 7：参与者对 Relatedly 和基线在多大程度上支持其文献综述流程的认同程度。对于每项陈述，我们报告李克特量表回答所占百分比，以及配对威尔科克森符号秩检验的结果，并列出 z 值和 p 值。

> "从逐篇阅读论文转变为阅读段落，需要一个适应过程；但一旦习惯了这种模式，以这种方式探索主题就会更容易。"（P13）

事后访谈中提到的其他大多数益处，都与实验室研究中发现的益处相呼应。此外，参与者报告说，他们会将 Relatedly 与其他应用搭配使用，例如：学术搜索引擎和 PDF 阅读器，用于仔细阅读论文；参考文献管理器（如 Zotero）、笔记应用（如 Notion）以及文档，用于附加笔记并整理论文以供日后使用。我们认为，这是一个积极的信号，表明*Relatedly* 也能够在较长时间内支持现实世界中的任务；而且，研究 1 的参与者在没有任何义务或报酬的情况下，自愿在研究结束后继续使用 *Relatedly*，这表明它至少为这五位学者带来了现实价值。

7 讨论

本文展示了如何利用先前的科学研究工作（即现有相关工作部分中的段落），为用户开展文献综述的发现与综合提供支架。受一项形成性研究的启发（该研究指出了这种方法面临的挑战），我们设计了一款名为 Relatedly 的新型系统，提供多种支架功能，例如自动生成描述性段落标题、对段落文本进行高亮和弱化，以便读者阅读重叠

PID	职位名称	使用原因 他们在……期间使用 Relatedly 研究工作流程	感兴趣的主题 兴趣	小时 使用 Relatedly/ 总时数 处理	# of 查询	# of 论文 精选
P07	研究 助理	查找并整理 相关论文，汇编成 带注释的阅读清单	智能手机无障碍功能 面向……人士的技术 运动障碍	4/4	10	30
P09	博士后 博士 研究人员	研究相关 工作以识别研究空白	认知与设计 反馈理论	25/36	14	102
P11	博士生	开始一个新项目，研究方向为 未知的研究主题	创意支持工具， 用于三维原型设计	10/10	4	98
P13	机器 学习 研究人员	检查未知项 撰写时的论文 关于已知主题的论文	多文档 摘要技术	5/5	3	51
P15	研究 科学家	检查未知 论文，以便撰写一篇 关于已知主题的论文	增强现实/虚拟现实技术的应用 于健康和教育领域	1.5/4	11	23

表 2：长期自愿使用 Relatedly 的参与者背景和使用情况。参与者自行报告了他们的职位名称、在研究工作流程中使用 Relatedly 的原因、工作时长、使用 Relatedly 的时长、查询数量，以及整理后准备阅读的相关论文数量。

和新颖的信息；重新排序段落以最大化子主题的广度，同时允许用户深入探索具体子主题。在此，我们将用户研究中的观察结果与理论联系起来，进一步阐释我们的研究发现。

教育心理学家 [1, 61] 发现，向学生提供同一主题的多种解释或不同表征，有助于他们克服任何特定解释或表征的局限，更好地理解复杂科学主题 [2] 并解决问题 [14]。在这种情况下，阅读由不同作者撰写、对主题的阐述可能略有不同的多个相似段落，或许能帮助用户更好地理解和综合相关信息。根据行为日志记录，我们确实发现，与使用基线相比，参与者使用 Relatedly 时浏览的段落显著更多；这或许可以解释为什么他们撰写的学习提纲质量显著高于使用基线时。

信息觅食理论指出，人们倾向于在不同信息斑块之间切换，以便在探索过程中最大限度地增加所获取信息的数量和广度 [56]。Relatedly 的概览视图和相似段落视图旨在帮助用户在探索从多篇论文中提取的相关工作段落时运用这一策略，这些段落包含重叠或不同的信息。在出声思维和任务后调查中，参与者表示，他们使用这两个视图，在探索不同子主题与深入阅读特定子主题的细节之间灵活切换；这或许可以解释为什么与基线相比，他们能够在学习提纲中写下显著更多的主题/子主题。

最后，参与者认为，与基线相比，Relatedly 对文献综述更有帮助。评估研究中的参与者更认同以下陈述：该系统帮助我“找到与主题相关的研究”“理解术语/概念之间的关系”“汇集来自多个来源和观点的信息”“确定阅读内容的优先级”以及“在文献综述流程中跟踪获取或阅读过的信息”，相比之下，他们对基线系统的认同度较低。总体而言，15 名参与者中有 13 名更愿意使用 Relatedly 而非基线来进行文献综述；研究结束后，其中 5 名参与者还采用了 Relatedly 来支持即将进行的论文投稿。考虑到参与者自愿持续使用该系统，既无义务，也无补偿，我们认为这表明 Relatedly 有望在参与者借助它完成自己的文献综述任务时带来实际益处。

我们认为，任何运用算法方法帮助用户管理注意力（推荐、搜索、摘要）的系统，在设计时都应意识到并考虑模型错误可能误导用户、隐性偏见以及回音室效应。开发 Relatedly 时，我们也致力于降低这些潜在风险。例如，我们谨慎把控 Relatedly 生成的章节标题的质量，并在自然语言处理（ROUGE）领域的标准自动评估技术之外，额外开展了人工评估。Relatedly 的若干用户界面功能也旨在应对这些风险。例如，进度条鼓励用户浏览更多论文和段落，而不是满足于已经探索过的内容。增加探索的益处有助于抵消所浏览论文分布可能发生的变化。此外，尽管大多数以往的推荐系统都帮助用户查找与已探索内容相似 的文档，Relatedly 则采取了不同做法：主动重新排列文档的顺序，并突出显示句子，以鼓励用户优先探索与已浏览内容最不相似 的信息。

Relatedly 的核心洞见是，当面对同一主题的多篇不同文本时，它会通过帮助读者优先探索新的不同信息，并降低冗余信息的优先级，为读者的探索提供支架。Relatedly 使用科学文本和相关工作章节展示了这一方法，不过，这种方法也可以

应用于其他领域，例如政策制定者在审阅政策文献以撰写政策简报时，律师在研究既往案件以发现模式、法律先例并进行辩护时，选民在多位政治人物的政纲或新闻文章中追踪重要议题时，或者程序员在不同的讨论论坛或 Jupyter 笔记本中寻找 bug 的解决方案时。我们设想一种增强阅读体验，帮助用户广泛了解不同观点或主题，优先选择阅读内容，并追踪自己在互联网上任何主题的信息来源中已读过的内容。我们将这些前景广阔的研究方向留待未来工作探索。

7.1 局限性与未来工作

许多设计决策都受到我们关注学者对科学主题进行文献综述这一点的影响。我们的一个假设是，科学家会在论文中撰写高质量的相关工作部分，相比于查阅单篇论文并进行综合，这些部分能够为用户带来更多益处。为此，我们从人机交互和自然语言处理领域的顶级会议中选取论文，以构建数据集。未来，如果希望通过这种方法将覆盖范围扩展到所有科学出版物，就需要进一步支持用户对提取相关工作段落的来源进行更细粒度的控制。例如，允许用户整理一组他们信任的会议或作者。或者，未来的研究可以进一步拓展用于评估写作质量的自然语言处理技术 [50, 66]，以自动评定相关工作部分的质量 [68]，并将其纳入排序算法。此外，进一步改进 Relatedly 的一个方向，是分析每条参考文献在其所在段落语境中的重要性。例如，可以使用诸如 [70]之类的自然语言处理技术，估计每条参考文献在段落中的影响力水平，从而使 Relatedly 能够减轻批量引用和顺带引用的影响 [26, 27, 37]。

在设计我们的用户研究时，我们还将引文图可视化工具作为一种额外的基线条件纳入了考虑。然而，文献综述任务可能难以开展研究 [40]，，因为由于其探索性，参与者可能会在心理上感到负担沉重且耗时较长 [40]。为了让研究保持现实性，我们希望让参与者投入较长时间的文献综述环节，同时将整个流程控制在 90 分钟以内，以避免疲劳。因此，我们选择了基线条件，该条件模拟了大多数用户最常采用的文献综述策略（即使用学术搜索引擎并逐篇阅读论文）。未来的工作可以在我们的研究基础上，将已有的可视化系统作为基线纳入其中，以便与 Relatedly 进行比较，从而帮助我们进一步理解 Relatedly 的成本与收益。

在形成性访谈中，我们主要采访了身为初级学者的博士生。由于 Relatedly 旨在通过对未知主题进行广泛概览并查找相关论文，帮助人们启动文献综述流程，因此我们重点了解了初级研究人员的需求。尽管形成性研究中的三名参与者是全职研究生阶段后的研究人员，但在形成性研究中，我们仍聚焦于初级研究人员，因为与资深学者相比，他们往往在文献综述流程方面面临更多挑战，也需要更多支持 [15, 19, 29, 67] 。资深学者更有可能依靠广泛的社交关系来支持论文推荐，并拥有更丰富的相邻领域知识可供借鉴 [15, 19]。现有研究系统通过根据社交信号推荐论文，支持资深研究人员进行文献检索，例如他们过去曾与哪些人合作或互动过 [34, 35]，，以及他们阅读或整理过的论文、作者、机构和发表场所 [36]。

为了避免剽窃等潜在的意外后果，Relatedly 的设计旨在突出思想来源，并通过将作者信息优先置于段落顶部，以及在用户复制参考文献时附加作者信息，来鼓励遵循正确的引用规范。

这一方法更广泛应用的一个潜在障碍，是许可问题以及对科学文献的访问。具体而言，并非所有学术论文都能由互联网上的任何人自由访问和搜索。另一方面，近年来推动开放科学 [48] 的趋势，以及 S2ORC 数据集、 [44], arXiv.org[22, 49], 和开放科学基金会[11]等举措，还有利用光学字符识别、GROBID、 [45], VILA [65], 等技术让较早发表的文章变得可访问，都指向一个充满希望的未来：学者能够更充分地利用彼此的既有研究成果，从而实现包括 Relatedly 在内的新技术和新型互动。

目前，Relatedly 面向学术用户进行设计。然而，一个有趣的未来方向是支持普通人，让探索科学信息变得更加容易。例如，如果某个人想要概览有关疫苗的科学文献，以决定是否接种疫苗，或者想要概览相关文献，将科学研究应用于初创产品。这一机会在于，看到不同论文的作者针对某一研究主题及彼此工作所阐述的不同观点，有可能避免普通人过度信任单一证据 [18]。在这种情况下，未来版本的 Relatedly 不仅可以将陌生术语链接到定义 [24]，或用通俗语言总结段落 [3]，，更重要的是，还可以呈现既有研究之间的一致与分歧及其不确定性程度，同时帮助用户对自己的学习建立信心和信任，这一点可能尤为重要 [18]。

知识工作和文献综述通常需要通过发起多个查询来探索多个主题 [19, 53]。参与者自愿使用 Relatedly 完成实际任务，其结果印证了这一点。虽然我们的用户评估实验室研究观察了 Relatedly 如何帮助参与者围绕单个主题和查询探索信息，但未来工作一个令人振奋的方向是研究用户如何在多个主题和查询之间转换探索方向。近期研究（如 [54, 55], ）帮助探索新领域的人们在缺乏领域专用语言和明确的信息目标时表述查询。我们将扩展 Relatedly 的方法留待未来工作，以更好地支持用户通过多个查询和主题转换来探索科学文献。

8 结论

本文探索了一种支持文献综述工作流程的新方法：我们不再专注于逐篇理解论文以掌握某个主题，而是让用户在许多相关工作

[11]开放科学基金会：https://www.cos.io/products/osf

章节中探索不同的子主题领域，这些章节段落从多篇论文中提取而来。其思路是，借助作者撰写相关工作章节时付出的科学研究工作，我们可以改进其他研究人员的文献综述流程。为了解决从不同文档中提取的段落可能涵盖相似或不同主题的问题，*Relatedly* 还提供阅读支持提示和信息增益跟踪功能，帮助用户阅读大量相关工作段落，从而更高效地全面了解不同主题。对比用户研究表明，Relatedly 的文献综述方法帮助学者以更广泛、更连贯且更有见地的方式综合主题信息。这可能是因为 Relatedly 的阅读支持功能为用户发现和浏览更多段落与论文提供了支架，从而帮助用户探索广阔且多面的信息空间。此外，使用 Relatedly 的参与者发现了更多领域专属术语，并且比起基线更倾向于使用它。我们认为，Relatedly 的方法让我们更进一步，能够利用网络上现有的信息与结构来支持知识探索与综合。

致谢

本项目获得美国国家科学基金会资助（项目编号 OIA-2033558）。作者谨向 Daniel S. Weld 和 Steven P. Dow 致谢，感谢他们富有见地的讨论和反馈。我们也感谢匿名评审提供的建设性意见。最后，若没有参与试点测试和用户研究的参与者，本研究便无法完成。

参考文献

[1] Shaaron Ainsworth。2006。DeFT：用于探讨借助多重表征进行学习的概念框架。《学习与教学》16, 3 (2006), 183–198。[2] Shaaron Ainsworth。2008。学习复杂科学概念时多重表征的教育价值。载于《可视化：科学教育中的理论与实践》。施普林格，191–208。[3] Tal August、Lucy Lu Wang、Jonathan Bragg、Marti A Hearst、Andrew Head 和 Kyle Lo。2022。Paper Plain：运用自然语言处理，让医疗研究论文更易于医疗保健消费者理解。*arXiv*预印本 *arXiv:2203.00130*（2022）。[4] Pia Borlund。2003。交互式信息检索评估模型：交互式信息检索系统评估框架。《信息研究》8, 3 (2003), 8–3。[5] Lutz Bornmann、Robin Haunschild 和 Rüdiger Mutz。2021。现代科学的增长率：一种用于建模既有和新兴文献数据库中出版物数量的潜在分段增长曲线方法。《人文与社会科学传播》8, 1 (2021), 1–15。[6] Christine Susan Bruce。1994。研究生早期的学位论文文献综述经历。《高等教育研究》19, 2 (1994), 217–229。[7] Isabel Cachola、Kyle Lo、Arman Cohan 和 Daniel Weld。2020。TLDR：科学文献的极限摘要。载于计算语言学协会研究发现：*EMNLP 2020*。计算语言学协会，线上，4766 4777 10 18653 1 2020 428 –。[8] https://doi.org/。/v/。findings-emnlp。Isabel Cachola、Kyle Lo、Arman Cohan 和 Daniel S Weld。TLDR：科学文献的极限摘要。*arXiv*预印本 *arXiv:2004.15011*（2020）。[9] Jaime Carbonell 和 Jade Goldstein。1998。MMR 的应用：基于多样性的重排序，用于重新排列文档并生成摘要。载于第*21*届 *ACM SIGIR* 信息检索研究与发展年度国际会议论文集。335–336。[10] Joseph Chee Chang、Aniket Kittur 和 Nathan Hahn。2016。Alloy：结合群体智慧与计算进行聚类。载于*2016*年 *CHI*计算机系统中的人为因素会议论文集。3180–3191。[11] Kathy Charmaz。2014。建构扎根理论。Sage。[12] Duen Horng Chau、Aniket Kittur、Jason I Hong 和 Christos Faloutsos。2011。Apolo：结合丰富的用户交互与机器学习，理解大型网络数据。载于*SIGCHI*会议计算机系统中的人为因素论文集。167–176。

[13] Arman Cohan、Sergey Feldman、Iz Beltagy、Doug Downey 和 Daniel S Weld。2020。SPECTER：利用引用信息增强型变换器进行文档级表示学习。arXiv预印本 arXiv:2004.07180（2020）。[14] Richard Cox 和 Paul Brna。1995。在问题解决中支持外部表征的使用：灵活学习环境的必要性。《人工智能教育杂志》6（1995），239–302。[15] Susan B Davidson 和 Juliana Freire。2008。溯源与科学工作流：挑战与机遇。载于《2008年ACM SIGMOD数据管理国际会议论文集》。1345–1350。[16] Andrew S Denney 和 Richard Tewksbury。2013。如何撰写文献综述。《刑事司法教育杂志》24，2（2013），218–234。[17] Jay DeYoung、Iz Beltagy、Madeleine van Zuylen、Bailey Kuehl 和 Lucy Lu Wang。2021。Ms2：医学研究的多文档摘要。arXiv预印本 arXiv:2104.06486（2021）。[18] Baruch Fischhoff。2013。科学传播的科学。《美国国家科学院院刊》110，增刊_3（2013），14033–14039。[19] Sarah Rose Fitzgerald。2017。高等教育领域学者的信息寻求。密歇根州立大学。[20] Santo Fortunato、Carl T Bergstrom、Katy Börner、James A Evans、Dirk Helbing、Staša Milojević、Alexander M Petersen、Filippo Radicchi、Roberta Sinatra、Brian Uzzi 等。2018。科学学。《科学》359，6379（2018），eaao0185。[21] Sebastian Gehrmann、Steven Layne 和 Franck Dernoncourt。2019。通过基于半马尔可夫条件随机场的神经网络章节标题生成，提升人类对文本的理解。载于《2019年北美计算语言学协会会议：人类语言技术论文集》第1卷（长短论文）。计算语言学协会，明尼阿波利斯，明尼苏达州，1677 1688 10 18653 1 19 1168 –。[22] 2011 20 476, 7359 2011 145 147 https://doi.org/ 。/v /N - Paul Ginsparg。 。arXiv： 。《自然》（），–。[23] Nathan Hahn、Joseph Chang、Ji Eun Kim 和 Aniket Kittur。2016。知识加速器：将整体性思维拆解为小块。载于《2016年CHI计算机系统中的人为因素会议论文集》。2258–2270。[24] Andrew Head、Kyle Lo、Dongyeop Kang、Raymond Fok、Sam Skjonsberg、Daniel S Weld 和 Marti A Hearst。2021。通过即时、位置敏感的术语和符号定义增强科学论文。载于《2021年CHI计算机系统中的人为因素会议论文集》。1–18。[25] Marti A Hearst。2006。用于信息探索的聚类与分面类别。《ACM通讯》49，4（2006），59–61。[26] Serge Horbach、Kaare Aagaard 和 Jesper W Schneider。2021。元研究：问题重重的引用实践如何扭曲科学。（2021）。[27] Serge PJM Horbach、Freek JW Oude Maatman、Willem Halffman 和 Wytske M Hepkema。2022。自动引文推荐工具会助长可疑引用。《研究评价》（2022）。[28] World Leaders in Research-Based User Experience。 [无日期。]大声思考：排名第一的可用性工具。https://www.nngroup.com/articles/thinking-aloud- the-1-usability-tool/ [29] Sharon Favaro Ince、Christopher Hoadley 和 Paul A Kirschner。2018。博士生学术工作流中的搜索实践研究。载于《2018年人类信息交互与检索会议论文集》。245–248。[30] Jill Jesson、Lydia Matheson 和 Fiona M Lacey。2011。撰写文献综述：传统与系统性方法。（2011）。[31] Ziwei Ji、Nayeon Lee、Rita Frieske、Tiezheng Yu、Dan Su、Yan Xu、Etsuko Ishii、Yejin Bang、Andrea Madotto 和 Pascale Fung。2022。自然语言生成中的幻觉综述。arXiv预印本 arXiv:2202.03629（2022）。[32] Arif E Jinha。2010。第五千万篇文章：对现存学术文章数量的估算。《学术出版》23，3（2010），258–263。[33] Hyeonsu B Kang、Joseph Chee Chang、Yongsung Kim 和 Aniket Kittur。2022。Threddy：用于个性化基于讨论串的科学文献探索与组织的交互式系统。arXiv预印本 arXiv:2208.03455（2022）。[34] Hyeonsu B Kang、Rafal Kocielnik、Andrew Head、Jiangjiang Yang、Matt Latzke、Aniket Kittur、Daniel S Weld、Doug Downey 和 Jonathan Bragg。2022。从认识的人到阅读的内容：利用隐式社交网络增强科学文献推荐。载于《CHI计算机系统中的人为因素会议》。1–23。[35] Hyeonsu B Kang、Nouran Soliman、Matt Latzke、Joseph Chee Chang 和 Jonathan Bragg。2023。ComLittee：利用个人选定作者委员会发现文献。载于《2023年CHI计算机系统中的人为因素会议论文集》（德国汉堡）（CHI '23）。计算机协会，美国纽约州纽约市。[36] Harmanpreet Kaur、Doug Downey、Amanpreet Singh、Evie Yu-Yen Cheng、Daniel Weld 和 Jonathan Bragg。2022。FeedLens：用于个性化知识图谱探索式搜索的多态透镜。载于《第35届ACM用户界面软件与技术年度研讨会论文集》。1–15。[37] Jerry S Kidd。1990。衡量引用实践。《美国信息科学学会杂志》41，3（1990），157–163。[38] Donald W King、Carol Tenopir、Songphan Choemprayong 和 Lei Wu。2009。五所美国大学教职人员的学术期刊信息寻求与阅读模式。《学术出版》22，2（2009），126–144。

[39] Jeffrey W Knopf。2006。撰写文献综述。《PS: Political Science & Politics》39，1（2006），127–132。[40] Bill Kules 和 Robert Capra。2009。为信息寻求支持系统的用户研究设计探索性搜索任务。载于《第9届ACM/IEEE-CS数字图书馆联合会议论文集》。419–420。[41] Clayton Lewis。1982。在认知界面设计中使用“大声思考”方法。IBM TJ Watson Research Center，约克敦高地，纽约州。[42] Mike Lewis、Yinhan Liu、Naman Goyal、Marjan Ghazvininejad、Abdelrahman Mohamed、Omer Levy、Veselin Stoyanov 和 Luke Zettlemoyer。2020。BART：用于自然语言生成、机器翻译和理解的去噪序列到序列预训练。载于《第58届计算语言学协会年会论文集》。计算语言学协会，线上，7871 7880 10 18653 1 2020 703 – 。[43] https://doi.org/ 。/v / .acl-main。2002 Chin-Yew Lin 和 Eduard Hovy。 。摘要的人工与自动评估。载于《ACL-02自动摘要研讨会论文集》。计算语言学协会，美国宾夕法尼亚州费城，45 51 10 3115 1118162 1118168 – 。[44] https://doi.org/ 。/ 。Kyle Lo、Lucy Lu Wang、Mark Neumann、Rodney Kinney 和 Daniel Weld。2020。S2ORC：Semantic Scholar开放研究语料库。载于《第58届计算语言学协会年会论文集》。计算语言学协会，线上，4969–4983。https://doi.org/10.18653/v1/2020. acl-main.447[45] Patrice Lopez。2009。GROBID：结合自动书目数据识别与学术出版物术语抽取。载于《数字图书馆理论与实践国际会议》。施普林格，473–474。[46] Kurt Luther、Nathan Hahn、Steven P Dow 和 Aniket Kittur。2015。Crowdlines：通过众包提纲支持多样化信息源的综合。载于第三届AAAI人类计算与众包会议。[47] Gary Marchionini。2006。探索式搜索：从查找到理解。《ACM通讯》49，4（2006），41–46。[48] Erin C McKiernan、Philip E Bourne、C Titus Brown、Stuart Buck、Amye Kenall、Jennifer Lin、Damon McDougall、Brian A Nosek、Karthik Ram、Courtney K Soderberg 等。2016。开放科学如何帮助研究人员取得成功。eLife 5（2016）。[49] Gerry McKiernan。2000。arXiv.org：洛斯阿拉莫斯国家实验室电子预印本服务器。《国际灰色文献期刊》（2000）。[50] Mohsen Mesgar 和 Michael Strube。2018。用于文本质量评估的神经局部连贯性模型。载于EMNLP。[51] James G Miller。1960。信息输入过载与精神病理学。《美国精神病学杂志》116，8（1960），695–704。[52] Keisuke Okamura。2019。重新审视跨学科性：研究影响力与动态性的证据。《帕尔格雷夫通讯》5，1（2019），1–9。[53] Srishti Palani、Zijian Ding、Stephen MacNeil 和 Steven P Dow。2021。“主动搜索”假设：搜索策略与创造性学习的关系。载于《2021年人类信息交互与检索会议论文集》。325–329。[54] Srishti Palani、Zijian Ding、Austin Nguyen、Andrew Chuang、Stephen MacNeil 和 Steven P Dow。2021。CoNotate：基于笔记提出查询，促进知识发现。载于《2021年CHI计算机系统中的人为因素会议论文集》。1–14。[55] Srishti Palani、Yingyi Zhou、Sheldon Zhu 和 Steven P Dow。2022。InterWeave：在上下文中呈现搜索建议，为信息搜索与综合提供支架。载于《第35届ACM用户界面软件与技术年度研讨会论文集》。1–16。[56] Peter Pirolli 和 Stuart Card。1995。信息访问环境中的信息觅食。载于《SIGCHI计算机系统中的人为因素会议论文集》。51–58。[57] Peter Pirolli 和 Stuart Card。2005。通过认知任务分析确定的意义建构过程及分析人员技术的杠杆点。载于《情报分析国际会议论文集》第5卷。美国弗吉尼亚州麦克莱恩，2–4。[58] Antoine Ponsard、Francisco Escalona 和 Tamara Munzner。2016。PaperQuest：支持文献综述的可视化工具。《2016年CHI计算机系统中的人为因素会议扩展摘要论文集》（2016）。[59] Colin Raffel、Noam Shazeer、Adam Roberts、Katherine Lee、Sharan Narang、Michael Matena、Yanqi Zhou、Wei Li、Peter J Liu 等。2020。利用统一的文本到文本转换器探索迁移学习的极限。《机器学习研究杂志》21，140（2020），1–67。[60] Stephen E Robertson 和 K Sparck Jones。1976。搜索词的相关性加权。《美国信息科学学会杂志》27，3（1976），129–146。[61] David Rosengrant、Eugenia Etkina 和 Alan Van Heuvelen。2007。多重表征近期研究概述。载于《AIP会议论文集》第883卷。美国物理联合会，149–152。[62] Stephen Rowland。2002。克服职业生活中的碎片化：学术发展的挑战。《高等教育季刊》56，1（2002），52–64。[63] Daniel M Russell、Mark J Stefik、Peter Pirolli 和 Stuart K Card。1993。意义建构的成本结构。载于《INTERACT'93与CHI'93计算系统中的人因会议论文集计算系统中的人因会议. 269–276. [64] Dafna Shahaf、Carlos Guestrin 和 Eric Horvitz。2012 年。《科学地铁图》。载于 第 *18* 届 *ACM SIGKDD* 知识发现与数据挖掘国际会议论文集。1122–1130。[65] Zejiang Shen、Kyle Lo、Lucy Lu Wang、Bailey Kuehl、Daniel S Weld 和 Doug Downey。2022 年。VILA：使用视觉布局组改进科学类 PDF 的结构化内容提取。《计算语言学协会汇刊》10（2022），376–392。[66] Kaveh Taghipour 和 Hwee Tou Ng。2016 年。一种自动作文评分的神经网络方法。载于 *2016* 年自然语言处理经验方法会议论文集。1882–1891。[67] Kornelia Tancheva、Gabriela Castro Gessner、Neely Tang、Erin Eldermire、Heather Furnas、Darcy Branchini、Gail Steinhart 和 Nancy Fried Foster。2016 年。一位（认真）研究人员的一天：展望研究图书馆的未来。《Ithaka》。网址：http://sr. ithaka. org (2016)。[68] Jaime Teevan。2014 年。学术论文的公式：相关工作。http://slowsearching.blogspot.com/2014/11/a-formula-for-academic-papers-related.html[69] Carol Tenopir、Rachel Volentine 和 Donald W King。2012 年。学者的论文和图书阅读模式：给出版商的研究发现。《学术出版》25，4（2012），279–291。[70] Marco Valenzuela、Vu A. Ha 和 Oren Etzioni。2015 年。识别有意义的引文。载于 *AAAI* 研讨会：学术大数据。[71] R Van Noorden。2014 年。全球科研产出每九年翻一番 [博客文章]。摘自 *nature. com*，网址：*http://blogs.nature.com/news/2014/05/global-scientific-output-doubles-every-nine-years.html*（2014）。[72] Richard Van Noorden 等。2015 年。用数字看跨学科研究。《自然》525，7569（2015），306–307。[73] Amy X Zhang、Jilin Chen、Wei Chai、Jinjun Xu、Lichan Hong 和 Ed Chi。2018 年。借助群体评估和改进聚类搜索结果。《ACM《交互式智能系统汇刊》 (TiiS)》8，2（2018），1–28。[74] Xiaolong Zhang、Yan Qu、C Lee Giles 和 Piyou Song。2008 年。CiteSense：支持研究文献的意义建构。载于 *SIGCHI* 计算系统人因会议论文集。677–680。

作者撰写的标题	模型生成的标题
无监督摘要生成	无监督抽象式摘要
边框发起的文本输入	智能手表上的文本输入
作为社交代理的机器人	面向社会表征的社交机器人设计
创客与创客空间	作为创造力场所的创客空间
社会文化因素与检查表效能	人工智能公平性中的文化张力
数据表提取与清洗	网页表格分类
双语词嵌入中的偏差	双语词嵌入

表3：并列展示的示例模型生成标题，适用于作者撰写的冗长且描述性强的段落标题。

作者撰写的标题	模型生成的标题
相关工作	机器翻译优化
清醒梦	清醒梦与虚拟现实
关于 Soylent	Soylent 作为产品与概念
引言	用于文本分割的主题建模
定义	视觉搜索与导航的延迟
结论性启示：数据管理的道德经济	
相关工作	经典关键词短语提取系统

表 4：示例模型 模型生成的标题：短小且泛泛的作者撰写标题与段落并列展示 e.

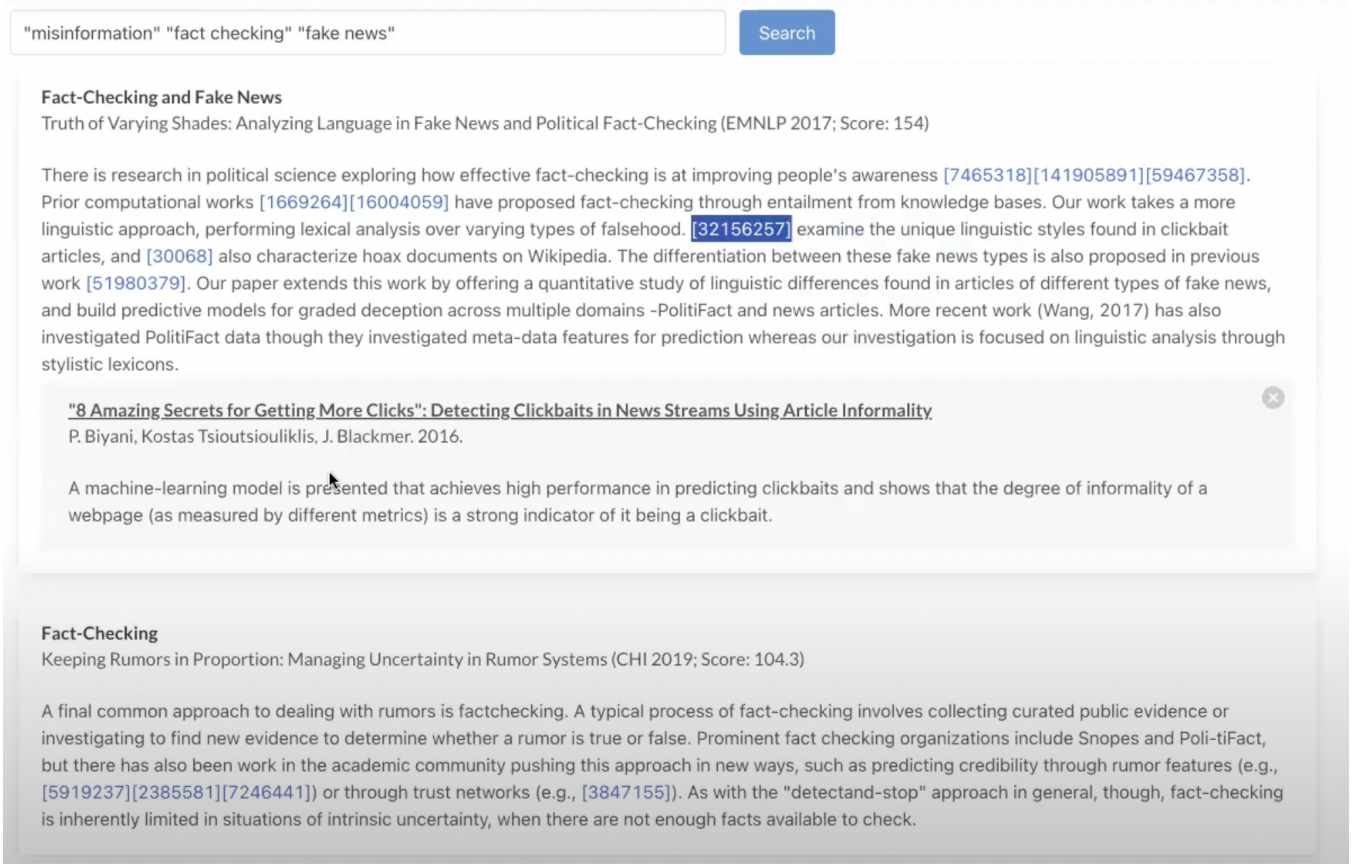

图 8：在形成性用户研究期间，参与者可以使用一个原型系统，在其中搜索主题，系统会从多篇论文的相关工作部分返回与主题相关的段落。他们可以点击参考文献（超链接语料库 ID）查看论文详情（包括链接到论文的论文标题、作者、摘要的 TLDR）

	Relatedly	基线	*z*	*p*
心理	14.52 ± 4.77	16.43 ± 4.35	-0.94	0.36
身体	13.33 ± 4.67	15.24 ± 4.29	-0.86	0.33
时间	12.39 ± 3.45	14.04 ± 2.49	-0.83	0.37
绩效	18.37 ± 4.34	12.00 ± 2.79	-0.67	0.49
努力程度	12.86 ± 4.79	14.05 ± 4.22	0.24	0.80
挫折感	6.83 ± 4.30	9.97 ± 6.34	1.59	0.11

表 5：使用 Relatedly 与基线时，任务工作负荷没有显著差异，这表明在感知工作负荷相近的情况下，性能有所提升。我们报告 NASA 任务负荷指数的平均得分及其标准差（作为不确定性的衡量），以及 Wilcoxon 符号秩检验的结果，包括 z 值和 p 值。可能值的范围为 1–20。