arXiv:2304.02623v1 [cs.CL] 5 Apr 2023

超越摘要：为真实世界的说明性写作任务设计人工智能支持

Zejiang Shen†、Tal August□、PaoSiangliulue□、Kyle Lo□、JonathanBragg□
Jeff Hammerbacher□、Doug Downey□、⋄、Joseph Chee Chang□、David Sontag†
†麻省理工学院，□艾伦人工智能研究所，⋄西北大学
{zjshen,dsontag}@mit.edu,{tala,paos,kylel,jbragg,jeffhammerbacher,dougd,josephc}@allenai.or
g

摘要
大型语言模型在设计和开发新的人工智能辅助写作支持工具方面带来了令人兴奋的新机遇和挑战。近期工作表明，利用这项新技术可以在许多场景中转变写作，例如创意写作中的构思、编辑支持和摘要。然而，人工智能支持的说明性写作——包括学者撰写文献综述或医生撰写病程记录等现实世界任务——相对研究不足。

在这篇立场论文中，我们认为，为说明性写作开发人工智能支持面临独特而令人振奋的研究挑战，并有望产生显著的现实影响。我们将说明性写作界定为基于证据的、生成知识的写作：它既包含对外部文档的摘要，也包含新的信息或知识。说明性写作可被视为作者围绕一组源文档进行意义建构过程的产物，而阅读、反思与写作之间的相互作用，也为设计人工智能支持带来了新的机遇。我们概述了人工智能支持设计的三个组成部分，并探讨了未来研究需要考虑的问题。

关键词
人工智能辅助写作、摘要、专家写作、增强型写作、说明性写作。

1 引言

大型语言模型（大语言模型） [4, 27, 32] 的出现给人工智能辅助写作的设计空间带来了巨大的变化。[1]大语言模型的语言理解能力和高质量文本生成能力，有望将高认知要求的写作任务半自动化，即根据简短的自然语言输入提示，帮助生成大纲，甚至生成篇幅较长且语法正确的段落。[2] 因此，研究界和商业界都越来越关注探索智能写作支持系统的新设计，包括支持创意故事写作、 [7, 21, 31],博客文章或电子邮件撰写、[3] 个人知识管理、[4] 等等。

尽管此前的研究探索了大语言模型的许多令人兴奋的应用，我们认为 **说明性写作**是现有人工智能增强写作研究中关注不足的一项任务。我们将说明性写作作品定义为总结事实并产生新知识或信息的文章。例如，这可能是研究人员收集并阅读多篇论文以撰写综述论文 [23],或医生审阅临床笔记以制定治疗方案 [18]。在这些情况下，作者不仅会总结源文档，还会添加来源中不存在的信息或提出新的见解，例如整理相关论文，或综合患者的症状和检查结果，以形成鉴别诊断并制定可能的治疗方案。

与不涉及外部文档的其他写作类型相比，例如创意故事或论证性写作 [21, 33],说明性写作要求作者理解来源、生成新的洞见，并在最终文章中准确引用和呈现提取的信息。阅读与写作之间的互动为开发人工智能支持带来了新的挑战，而对于如何融入最新一代语言模型，这一设计在很大程度上仍未得到探索。尽管摘要通常是必要的组成部分，但说明性写作在目标上也不同于文档摘要任务：它旨在带来来源中不存在的新信息。

说明性写作可以被视为一个**意义建构**过程 [29],，其中涉及不同类型的子任务：通常，作者首先通过迭代探索和阅读多份相关文档来识别并提取关键证据，然后将证据组织成有用的框架，并进一步综合成连贯的写作，以传达新的知识或信息 [24]。因此，人工智能在写作过程中所扮演的角色可能有所不同，我们认为，设计不同类型和层级的人工智能支持非常重要，从而在最大限度地帮助作者的同时，尽可能减少对其观点的影响或塑造。在一种方法中，在写作的早期阶段，写作者会发起并推动工作，而人工智能应仅提供有限的支持功能。随着想法逐渐成形，作者对写作内容有了更好的把握，人工智能可以承担更多责任，由作者“监督”模型的工作。例如，在开始撰写综述论文时，作者首先提出查询，以查找相关论文，而人工智能则帮助执行搜索并发现相关文档；作者阅读检索到的论文并形成写作想法后，人工智能可以根据作者的想法帮助生成写作文本，写作者只需校对模型输出即可。人工智能发现有助于用户更好地学习，而编辑支持能让写作更加高效和完善；最重要的是，写作者能够完全掌控思考过程以及所产出文章中包含的想法。

说明性写作在许多现实世界任务中经常发生，我们认为现实世界说明性写作 是人工智能增强型写作支持的一个高价值开放问题，该领域存在许多挑战和机遇。许多现实的说明性写作

[1]在本研究的其余部分，除非另有说明，AI 辅助写作是指使用大语言模型来支持写作。[2] https://platform.openai.com/examples[3]例如，copy.ai 等商业工具，https://copy.ai 以及 Lex，https://lex.page/。[4]包括 mem 等应用，https://get.mem.ai 以及 Notion-AI, https://www.notion.so/。

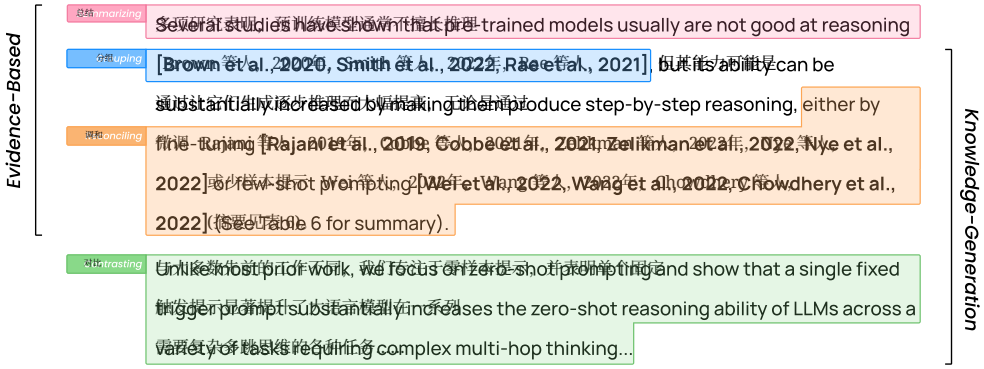


图1：剖析一篇说明性写作文本。在此示例中，这段写作是论文相关工作部分中的一段 [20]。其中包含相关先前研究的摘要，但也选择将几篇论文归为一组，因为相似的概念，或在共同框架下调和不同的研究，最后，它将当前论文与先前研究进行对比。这段写作**基于现有证据**，但也**呈现了关于现有研究之间关系的**新知识，以及这些新颖性；这些都是作者在写作过程中进行意义建构后形成的最终成果。**意义建构**过程。

任务需要领域专家，例如描述诉讼案件中的关键事件 [30], 解释科学概念或思想 [19], 以及介绍患者的病情和治疗方案 [1]。成功应用于这些任务的增强型写作系统不仅能够减少完成写作所需的昂贵专家工时，还能提高输出质量（例如，Bell 等人 [3] 报告称，每 5 名患者中就有 1 名会在医生或护士撰写的临床笔记中发现错误）。通常，为此类现实世界任务提供支持也能让我们利用现有的丰富示例数据资源库和成熟的评估协议，从而进一步推动未来 AI增强型写作系统的发展。在接下来的章节中，我们将正式定义说明性写作任务，并概述相关设计组件，这些组件旨在辅助写作过程，而不仅仅是辅助最终文章的完成。

2 界定说明性写作

与其他写作任务相比，说明性写作有两个独特特点：如图 *1*所示，它是证据驱动的，也是生成知识的。形式上，给定一组源文档和一系列写作目标，说明性写作旨在撰写一篇文章，综合来源中的信息，并根据目标产生新信息。作者综合来源的方式多种多样，包括但不限于对多条相似或不同（甚至相互矛盾）陈述进行筛选、分组、对比或调和；作者的综合也会带来源文档中未包含的信息。例如，医生可能会从临床笔记中筛选并分组一组相关的观察结果，并结合一种可能的病症，以实现制定治疗方案的目标。写作目标既指导阅读过程，也指导综合过程；在写作过程中产生新的见解后，写作目标也可以更新。与摘要任务不同，在这里，最终写作成果中不必纳入所有源文档：事实上，不纳入某份文档这一选择本身也会在写作成果中构成新信息（即表明该文档对于实现目标的相对重要性）。

3 设计用于说明性写作的人工智能支持

为了以最佳方式引入大语言模型来辅助说明性写作，我们认为需要包含三个组成部分：1）支持阅读和证据搜寻，以便正确、高效地理解来源；2）辅助信息综合以及新知识和创意的产出；3）促进文本撰写，以传达相关证据和洞见。说明性写作作品旨在带来新知识或新视角，但作者需要付出大量努力才能理解来源，并将思考转化为写作。通过减轻阅读过程中信息提取和理解所带来的认知负荷与交互成本，可以帮助作者专注于推理和创意生成 [16, 23]。另一方面，尽管大语言模型在文档理解和文本生成方面表现出色，但目前仍存在幻觉问题 [15]，且推理 [4, 32] 和长文本生成能力有限。因此，使用大语言模型一次性生成整篇作品再进行后期编辑并非最优方案 [6]。在写作的不同阶段提供不同人工智能支持的模块化设计最有帮助，以下将详细介绍各个组成部分。

增强阅读与证据收集。阅读相关文档以收集证据，是说明性写作的关键早期步骤。尽管已有大量关于文档发现和理解的研究工作（例如，针对科学文档 [8, 9, 14, 17, 25]），但支持为写作而阅读的人工智能工具相对稀缺 [12, 16, 26]。专家通常需要投入大量人工精力，阅读许多源文档，以识别关键的相关信息，帮助他们综合新知识。

这被称为“建立工作记忆” [2, 13], , 这是写作认知过程中的关键步骤。促进阅读的一种现有方法是使用语言模型为长文档自动生成摘要 [4, 34], , 但它们也容易产生幻觉 [15]。更重要的是，Ziegler等人 [35] 表明，近期的指令微调模型在执行摘要任务时仅仅是“聪明的复印机”，而过度信任模型输出可能会使作者观点产生偏差，并可能导致作者收集不准确的证据。

因此，作者阅读并理解来源仍起着核心作用，为说明性写作而阅读进行设计应侧重于对阅读的增强以及从源文档中提取关键信息 [16], , 但同时提供支持以验证提取的证据。例如，系统可以从一组源文档中自动生成与作者目标相关的关键事实列表，并通过将每项事实链接回原始源文档的相关部分来引导作者检查并核实事实的正确性。

支持信息综合。作为阅读与写作之间的重要中间步骤，作者通常需要花费很长时间来检查并推理所收集的证据，并产生待写入的新想法，例如，找出论文之间的异同并在共同框架下调和它们，如图1所示。综合跨源文档收集的大量信息可能具有挑战性，已开发出不同的界面和技术来支持此步骤 [5, 12, 26]。语言模型可通过例如基于文本表示的自动分组 [28] 或启用语义搜索来提高这些方法的效率和准确性。除了更好地组织证据以反映作者的心智模型外，系统还可以在综合过程中利用大语言模型来提供灵感 [11]。一个例子是提出证据之间的联系并建议相关/不同的主题以供发现，不过同样重要的是提供机制，使作者能够用相关证据验证并扎根模型生成的想法。人工智能模型的参与并非旨在替代作者的思考与沉思，而是增强新想法的产生，并发现否则看不见的联系或发现。

促进文本撰写。将想法转化为语言被认为在认知上与其他组成部分同样（甚至更加）要求高且具有挑战性：正如 Flower 和 Hayes [10] 所说，它需要“应对书面英语的所有特殊要求”，并且“这一过程可能会干扰……规划”。最新的语言模型已展现出生成流畅且高质量文本的强大能力，一些研究表明，在某些摘要任务中，其生成内容甚至可与自由撰稿人的作品相媲美 [34]。将语言模型纳入文本撰写支持颇具前景，但需要考虑若干设计选择，以降低生成错误和影响作者观点的风险 [22]。例如，在此情境下，语言模型可能需要具备证据感知能力并保持明确性——只有在作者调用时才应触发生成，同时还需要提供相关证据以及一条关于写作意图的简短提示。一次模型生成的内容可以相对简短，例如一到两句话，由作者调用它在多轮中。与一次性生成长补全（例如整段）相比，我们推测这可以降低错误和有偏见生成的可能性，并降低作者验证的成本。

4 结论

说明性写作是一种证据驱动且知识生成的写作体裁，发生在许多现实世界环境中。在本文中，我们论证说明性写作的独特特征为设计人工智能支持开辟了新机会。我们勾勒了设计的组成部分，并强调了未来实施的考虑因素和挑战。

致谢

我们感谢匿名审稿人提出的富有洞察力的评论和建议。我们感谢 Oren Etzioni、Arvind Satyanarayan、Dennis Wei、Prasanna Sattigeri、SubhroDas、BarbaraLam、LaurenYu 和 Ruochen Zhang 提供的有益讨论与建议。

参考文献

[1] Griffin Adams、Emily Alsentzer、Mert Ketenci、Jason Zucker 和 Noémie El-hadad。2021年。摘要中有什么？为推进住院病程摘要工作奠定基础。载于2021年计算语言学协会北美分会会议论文集：人类语言技术。计算语言学协会，线上，4794 4811 10 18653 1 2021 382 – 。https://doi.org/ 。/v / . [2] 1992 255, 5044 naacl-main。Alan Baddeley。工作记忆。Science（1992），556–559。[3] Sigall K Bell、Tom Delbanco、Joann G Elmore、Patricia S Fitzgerald、Alan Fossa、Kendall Harcourt、Suzanne G Leveille、Thomas H Payne、Rebecca A Stametz、Jan Walker 等。2020年。电子健康记录门诊记录中患者报告错误的频率和类型。JAMA Network Open 3, 6（2020），e205867–e205867。[4] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell、Sandhini Agarwal、Ariel Herbert-Voss、Gretchen Krueger、Tom Henighan、Rewon Child、Aditya Ramesh、Daniel Ziegler、Jeffrey Wu、Clemens Winter、Chris Hesse、Mark Chen、Eric Sigler、Mateusz Litwin、Scott Gray、Benjamin Chess、Jack Clark、Christopher Berner、Sam McCandlish、Alec Radford、Ilya Sutskever 和 Dario Amodei。2020年。语言模型是少样本学习者。载于《神经信息处理系统进展》，H. Larochelle、M. Ranzato、R. Hadsell、M.F. Balcan 和 H. Lin（编），第33卷。Curran Associates, Inc.，1877–1901。[5] Arie Cattan、Sophie Johnson、Daniel Weld、Ido Dagan、Iz Beltagy、Doug Downey 和 Tom Hope。2021年。Scico：面向科学概念的层次化跨文档共指。arXiv预印本 arXiv:2104.08809（2021）。[6] Ruijia Cheng、Alison Smith-Renner、Ke Zhang、Joel R Tetreault 和 Alejandro Jaimes。2022年。绘制文本摘要中人机交互的设计空间。arXiv预印本 arXiv:2206.14863（2022）。[7] John Joon Young Chung、Wooseok Kim、Kang Min Yoo、Hwaran Lee、Eytan Adar 和 Minsuk Chang。2022年。TaleBrush：使用生成式预训练语言模型构思故事。载于2022年计算机系统中的人为因素会议论文集。1–19。[8] Arman Cohan、Sergey Feldman、Iz Beltagy、Doug Downey 和 Daniel S. Weld。2020年。SPECTER：使用引文感知Transformer进行文档级表示学习。载于计算语言学协会年会。[9] Joseph Chee Chang 等。2022年。CiteSee：为科学论文中的引文补充持久且个性化的历史背景。ArXiv abs/2022.99999（2022）。[10] Linda Flower 和 John R Hayes。1981年。写作的认知过程理论。《大学写作与交流》32, 4（1981），365–387。[11] K. Gero、Vivian Liu 和 Lydia B. Chilton。2021年。Sparks：利用语言模型激发科学写作灵感。交互系统设计会议（2021）。[12] Han L. Han、Junhang Yu、Raphael Bournet、Alexandre Ciorascu、Wendy E. Mackay 和 Michel Beaudouin-Lafon。2022年。Passages：跨文档与文本交互。2022年计算机系统中的人为因素会议论文集（2022）。[13] John R Hayes。2000年。理解写作中认知与情感的新框架。《写作视角：研究、理论与实践》（2000），6–44。[14] Andrew Head、Kyle Lo、Dongyeop Kang、Raymond Fok、Sam Skjonsberg、Daniel S. Weld 和 Marti A. Hearst。2020年。为科学论文补充即时、

术语与符号的位置敏感定义。 *2021*年计算机系统中人为因素国际会议论文集 (2020年).[15] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, AndreaMadotto, andPascale Fung.2022年。自然语言生成中的幻觉综述。 《计算机调查》 (2022年). [16] Hyeonsu B Kang,Joseph Chee Chang,Yongsung Kim,and Aniket Kittur.2022年。Threddy：通过个人选举作者委员会进行科学文献的个性化线程式探索与组织。 *arXiv*预印本 *arXiv:2208.03455* (2022年). [17] Hyeonsu B Kang, Nouran Soliman, Matt Latzke, Joseph Chee Chang, and Joseph Chee Chang. 2023年。ComLittee：利用个人选举作者委员会进行文献发现。 *ArXiv* abs/2302.06780 (2023年). [18] Amy JHKindand MaureenASmith.2008年。急性护理向亚急性护理转变过程中强制出院摘要组成部分的记录。患者安全进展：新方向与替代方法（第*2*卷：文化与再设计） (2008年). [19] Daniel King,Zejiang Shen, Nishant Subramani, Daniel S. Weld, IzBeltagy,andDoug Downey. 2022年。不要说你不知道的事：通过约束集束搜索提高抽象式摘要的一致性。 *ArXiv* abs/2203.08436 (2022年).[20] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, YutakaMatsuo, andYusuke Iwasawa. 2022年。大型语言模型是零样本推理器。 *arXiv*预印本 *arXiv:2205.11916* (2022年). [21] MinaLee,Percy Liang, and Qian Yang.2022年。Coauthor：设计用于探索语言模型能力的人机协作写作数据集。在 *2022*年计算机系统中人为因素国际会议论文集中。1–19.[22] Piotr Mirowski, Kory W Mathewson, JaylenPittman, and RichardEvans. 2022年。使用语言模型共同创作电影剧本和戏剧剧本：行业专业人士的评估。 *arXiv*预印本 *arXiv:2209.14958* (2022年).[23] Srishti Palani, Aakanksha Naik, Doug Downey, Amy X Zhang, Jonathan Bragg, and JosephChee Chang. 2023年。Relatedly：利用现有相关工作章节为文献综述提供支架。 *arXiv* 预印本 *arXiv:2302.06754* (2023 年). [24] Peter Pirolli.2007 年。通过认知任务分析确定的意义建构过程与分析师技术的杠杆点。 [25] Napol Rachatasumrit, Jonathan Bragg, Amy X.Zhang, andDanielS. Weld.2022年。CiteRead：将局部引文语境集成到科学论文阅读中。 第*27*届智能用户界面国际会议 (2022年). [26] NapolRachatasumrit, Gonzalo A. Ramos, Jina Suh,RachelNg, and Christopher Meek.2021年。ForSense：通过意义建构集成和机器研究支持加速在线研究。 第*26*届智能用户界面国际会议 (2021年). [27] Jack W. Rae, Sebastian Borgeaud, Trevor Cai,Katie Millican, Jordan Hoffmann, Francis Song,John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John F. J. Mellor,Irina Higgins, Antonia Creswell, Nathan McAleese,Amy Wu, Erich Elsen, SiddhantM. Jayakumar,Elena Buchatskaya, David Budden,Esme Sutherland, Karen Simonyan, Michela Paganini, L. Sifre,Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsi mpoukelli, N. K. Grigorev, Doug Fritz, Thibault Sottiaux,Mantas Pajarskas, Tobias Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin,Aidan Clark,Diego de LasCasas, Aurelia Guy, Chris Jones, James Bradbury, Matthew G. Johnson, Blake A. Hechtman, Laura Weidinger,Iason Gabriel, William S. Isaac,Edward Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol Vinyals, Kareem W. Ayoub, Jeff Stanway, L. L. Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2021年。扩展语言模型：Gopher训练的方法、分析与洞见。 *ArXiv* abs/2112.11446 (2021年). [28] NilsReimers and Iryna Gurevych.2019年。Sentence-BERT：使用孪生BERT网络的句子嵌入。 *arXiv*预印本 *arXiv:1908.10084* (2019年). [29] Daniel M Russell, Mark J Stefik, PeterPirolli, and Stuart K Card. 1993年。意义建构的成本结构。在 *INTERACT'93*与*CHI'93*计算机系统中的人为因素会议论文集中。269–276.[30] Zejiang Shen, Kyle Lo, Lauren Yu, Nathan Dahlberg, MargoSchlanger, andDougDowney. 2022年。Multi-LexSum：多粒度的民权诉讼现实世界摘要。 *arXiv*预印本 *arXiv:2206.10883* (2022年). [31] Nikhil Singh, Guiller mo Bernal, DariaSavchenko,andElenaL Glassman. 2022年。在哪里藏一头被盗的大象：多模态机器智能在创意写作中的飞跃。 《*ACM*计算机与人类交互汇刊》 (2022年). [32] Shaden Smith,Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Anand Korthikanti, Elton Zhang, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. 2022年。使用DeepSpeed和Megatron训练Megatron-Turing NLG 530B这一大型生成式语言模型。 *ArXiv* abs/2201.11990 (2022年)。

[33] Kevin Yang、Nanyun Peng、Yuandong Tian 和 Dan Klein。2022年。Re3：通过递归重新提示与修订生成更长的故事。*arXiv* 预印本 *arXiv:2210.06774*（2022年）。[34] Tianyi Zhang、Faisal Ladhak、Esin Durmus、Percy Liang、Kathleen McKeown 和 Tatsunori B. Hashimoto。2023年。大型语言模型新闻摘要基准测试。 *CoRR*（2023年）。arXiv：2301。13848 [cs.CL] http://arxiv.org/ abs/2301.13848v1[35] Daniel M Ziegler、Nisan Stiennon、Jeffrey Wu、Tom B Brown、Alec Radford、Dario Amodei、Paul Christiano 和 Geoffrey Irving。2019年。基于人类偏好对语言模型进行微调。*arXiv* 预印本 *arXiv:1909.08593*（2019年）。