arXiv:2304.02822v1 [cs.HC] 6 Apr 2023

运用行动七阶段探讨智能写作助手的可用性

Avinash Bhat
麦吉尔大学
加拿大

Disha Shrivastava
Mila，蒙特利尔大学
加拿大

Jin L.C. Guo
麦吉尔大学
加拿大

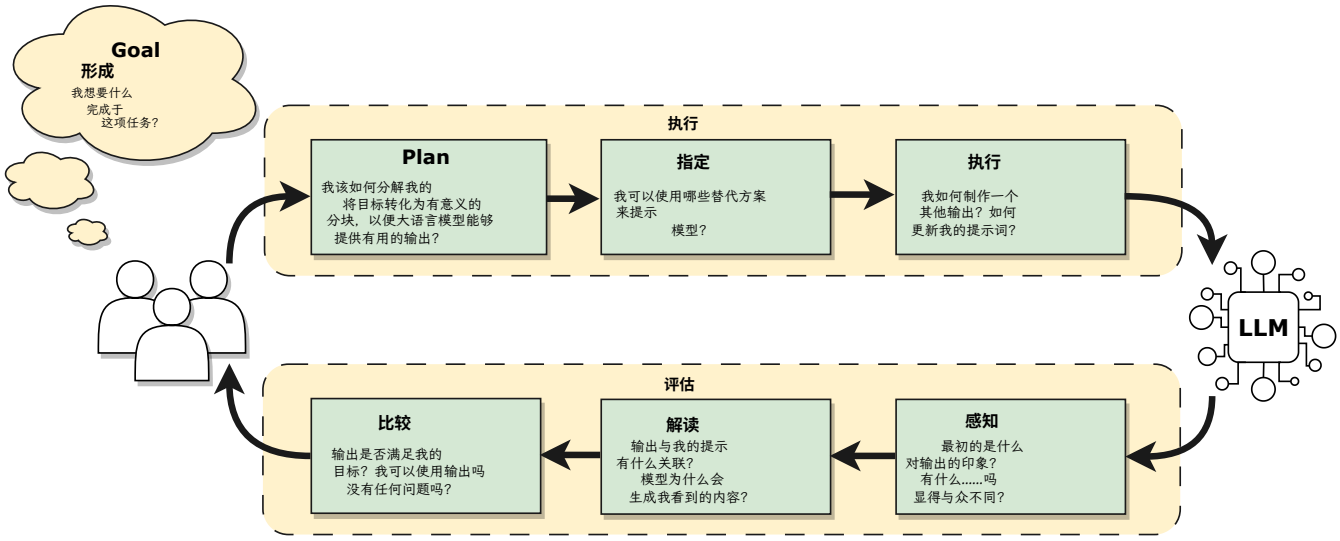

图 1：在与大语言模型交互的语境下考察的诺曼的行动七阶段示意图。

摘要

尽管大语言模型（LLM）作为写作助手潜力巨大，但它们仍受困于模型输出的连贯性与流畅性、可信度、生成内容的所有权以及模型性能的可预测性等问题，从而限制了其实用性。在这篇立场论文中，我们提出采用诺曼的行动七阶段作为框架，以探讨智能写作助手的交互设计。我们通过软件教程编写的示例，说明该框架在写作任务中的适用性。本文还探讨了该框架如何作为综合有关基于大语言模型工具的交互设计研究的工具，并介绍了支持行动各阶段的工具示例。最后，我们简要概述了该框架在人与大语言模型交互研究中的潜力。

1 引言

智能写作助手已针对各种写作目标和活动得到广泛研究 [8]。写作助-手的近期进展主要围绕大语言模型（大语言模型）展开 [22],，人类可以利用这些模型，根据提示中提供的意图生成内容。ChatGPT等大语言模型的显著进展[1] 及其在日常产品中的应用[2] 凸显了它们作为写作助手的潜力。然而，人机交互与此类助手协作会暴露出与其可用性相关的主要局限，例如模型输出的连贯性和流畅性 [9, 21],可信度 [10], ，以及生成内容的所有权 [1] 和模型性能的可预测性 [9, 12]。这些问题往往导致用户无法有效使用这些工具来实现其写作目标，有时甚至完全弃用它们。

尽管此前的研究 [4, 8, 14] 在一定程度上探讨了写作助手的交互方面，但尚无专门研究致力于实现端到端的写作目标，也未从可用性的角度审视其交互。我们从这些研究和现有设计文献中汲取灵感，聚焦人的行动，探究由大语言模型支持的智能写作助手中的交互设计。我们进一步提出采用诺曼的行动七阶段 [17] 作为框架，指导大语言模型支持的智能写作助手设计，并探讨其对可用性的影响。

2 行动的七个阶段

诺曼的行动七阶段是一种循环认知模型，通常用于理解用户的心理过程及其相应的身体动作，主要用于指导系统的交互设计。如图 1 所示，行动七阶段包括 (a) 形成目标、(b) 计划、(c) 明确、(d) 执行、(e) 感知、(f)解释和 (g) 比较。计划、明确和执行阶段构成执行阶段；感知、解释

[1]https://openai.com/blog/chatgpt[2]
https://openai.com/blog/chatgpt-plugins

和比较阶段构成交互的评估阶段。用户与系统的交互取决于用户基于先前认知形成的心智模型。我们认为，这一框架为设计和评估界面提供了契机，使界面能够支持用户在与大语言模型交互的不同阶段采取细粒度操作。我们提出，有效的基于大语言模型的写作助手设计必须回答与各个阶段相关的问题，从而指导设计并为用户提供必要的功能。

为了进一步阐释我们的想法，我们举例说明。这个例子主要受到我们最初尝试利用 OpenAI的Codex [2] 编写软件教程的启发。在典型的交互中，用户首先会形成一个想要达成的初步目标，例如，编写一篇介绍如何使用 matplotlib 绘制数据点的教程。接下来，用户会规划 目标，将其划分为相关部分，以便据此向写作助手提问。例如，可以将总体目标拆分为几个步骤：先撰写教程章节，例如提供适用于不同环境的库安装命令；接着生成并解释代码片段；最后提高教程的可读性。这里的每个步骤也可以视为一个子目标，其范围较小，并且可以经历行动框架的多轮迭代。当用户向写作助手发出提示时，通常会在写作助手界面中明确说明 需求，随后执行 需求，例如：“给定 Python 列表中的数据点，编写一段使用 matplotlib 绘制散点图的代码，并解释这段代码。”明确说明 阶段可以设置机制，向模型推荐不同的提示语；而执行 阶段则可以提供不同的界面功能，用于编辑和更新提示语。用户对任务和领域的了解，以及他们已有的概念模型，都会影响执行阶段。写作助手返回输出后，用户会根据自身的知识和专业水平感知 并解读 输出，同时更新已有的心智模型。例如，熟悉 matplotlib 的用户或许能更好地察觉生成代码中不寻常的内容或错误。用户可能还需要将输出与不同环境中的资源进行比较，例如，在 IDE 中执行生成的代码片段，或运行现有的单元测试。针对各个阶段提出相关问题，有助于找出关键交互，并指导写作助手的设计，以支持教程编写任务。

3 讨论

诺曼的七阶段行动框架补充了写作认知过程理论 [6]，正如Gero等人 [8] 在其关于写作助手设计空间的研究中所讨论的那样。写作认知过程理论指出，写作过程中会发生若干关键过程，例如构思、组织内容并设定写作目标，以及修改已写内容。可以在这些认知过程的每个阶段引入行动阶段，将大语言模型作为辅助写作者的补充型代理。鉴于大语言模型在不同写作任务中表现出的有效性，在研究写作工具的设计空间时，或许有必要将交互纳入考量维度。值得注意的是，Lee等人 [14] 近期开展的一项研究根据大语言模型在人-大语言模型交互环境中的表现对其进行了评估，并强调了研究此类交互的重要性，以便更好地评估模型。受这些研究启发，我们提议引入行动七阶段框架，以拓展关于基于大语言模型的工具所需能力与交互，以及相应影响的讨论。

行动七阶段框架有助于综合现有关于基于大语言模型的工具所采用交互策略的研究，显示出对执行阶段（或许并不理想）的侧重。诸如提示词串联 [20]，（涉及如何设计提示词）等现有策略，可以被视为支持规划阶段和规范阶段的工作。另一个支持执行阶段的工具示例是 PromptMaker[13]，它为设计提示词提供了替代界面。

具体而言，在基于大语言模型的写作助手的交互中，Wordcraft 中的元提示策略 [22] 为用户提供替代提示词，协助完成规范阶段。TaleBrush [5] 向模型引入了替代性的草图输入，以引导内容生成，展示了规范阶段和执行阶段中的交互可能性。在评估阶段，PromptChainer [19]将提示词设计的每一步输出进行可视化，通过在用户规范与大语言模型输出之间提供有效映射，实现更好的交互循环。

从七阶段行动框架的视角审视人类与大语言模型的交互，有望帮助我们更好地管理并应对相关挑战。例如，将大语言模型用作写作助手的后端存在风险 [18]，包括基于社会刻板印象的歧视以及针对性仇恨言论等公平性问题。虽然可以采用一些策略来识别 [16] 并限制[3] 某些滥用情形，但鉴于交互具有开放性，要完全对其进行审核将极具挑战，甚至不可能 [7]。在某些情况下，考虑到创意表达与令人反感的输出之间界限微妙，审核甚至可能不可接受 [12]。了解当前人类与大语言模型交互的心智模型，并将其分解为具有特定动机的各个阶段，有望助力内容审核。例如，采用七阶段行动框架的助手可以利用规划 阶段来界定并审核输出范围，利用具体说明 阶段来监督提示词，确保用户不会要求生成任何不当内容。解读 和比较 阶段可用于核实输出的相关性并减少任何不当内容。

我们预见到，使用该框架所面临的一项挑战是，七个阶段之间的区别并不总是显而易见，尤其是在处理写作等复杂任务时。例如，输出的感知与解释阶段之间的界限可能难以辨别，因为这些过程往往同时发生并相互作用。在所有情况下，严格遵循该框架可能都不可行。为克服这一局限，设计师可以采取灵活的方法，兼顾任务的独特特点和用户需求，并针对具体任务进行专门设计。根据使用场景，设计师可以确定需要重点关注的阶段，或以不同方式实现各个阶段。

例如，处理软件教程和研究写作的助手在比较阶段可能都试图

[3]https://platform.openai.com/docs/guides/moderation

实现确保输出准确这一相同的设计目标，但前一种使用场景可以通过运行代码或教程来实现这一目标，而后一种使用场景则可以通过提供指向数字图书馆的链接来获取可验证的信息。任何细微差异和特定任务的要求，都可以通过以用户为中心的设计方法（如用户访谈或参与式设计）加以识别，设计师则可以利用这些信息创建符合目标的解决方案。

结论。我们认为，采用诺曼的行动七阶段作为探索用户使用基于LLM的写作助手时行为的框架，可以为实现和设计涵盖目标形成、执行和评估阶段的细粒度交互提供有价值的结构。根据该框架所提出的交互设计维度，分析这些工具及其功能，可以应对基于LLM 的写作工具设计中的具体可用性问题。更进一步地说，这些维度表明，人类与大语言模型的交互研究中存在一些目前尚未得到充分体现的潜在方向，例如与人类偏好的对齐 [3], 有效的提示词设计 [15] 以及模型输出的可解释性和可阐释性 [11]。

参考文献

[1] Daer Gmann、Ning F. Ma 和 Dongwook Yoon。2022年。《从工具到伙伴：故事作者希望 AI 写作者尊重他们的个人价值观与写作策略》。载于交互系统设计会议（线上活动，澳大利亚）（*DIS '22*）。计算机协会，纽约州纽约市，美国，1209–1227。https://doi.org/10.1145/3532106.3533506[2] Mark Chen、Jerry Tworek、Heewoo Jun、Qiming Yuan、Henrique Ponde de Oliveira Pinto、Jared Kaplan、Harri Edwards、Yuri Burda、Nicholas Joseph、Greg Brockman、Alex Ray、Raul Puri、Gretchen Krueger、Michael Petrov、Heidy Khlaaf、Girish Sastry、Pamela Mishkin、Brooke Chan、Scott Gray、Nick Ryder、Mikhail Pavlov、Alethea Power、Lukasz Kaiser、Mohammad Bavarian、Clemens Winter、Philippe Tillet、Felipe Petroski Such、Dave Cummings、Matthias Plappert、Fotios Chantzis、Elizabeth Barnes、Ariel Herbert-Voss、William Hebgen Guss、Alex Nichol、Alex Paino、Nikolas Tezak、Ji e Tang、Igor Babuschkin、Suchir Balaji、Shantanu Jain、William Saunders、Christopher Hesse、Andrew N. Carr、Jan Leike、Josh Achiam、Vedant Misra、Evan Morikawa、Alec Radford、Matthew Knight、Miles Brundage、Mira Murati、Katie Mayer、Peter Welinder、Bob McGrew、Dario Amodei、Sam McCandlish、Ilya Sutskever 和 Wojciech Zaremba。2021年。《在代码上训练的大语言模型的评估》。https://doi.org/ 10 48550 2107 03374 。/ARXIV。。[3] Paul Christiano、Jan Leike、Tom B. Brown、Miljan Martic、Shane Legg 和 Dario Amodei。2017年。《基于人类偏好的深度强化学习》。https: //doi.org/ 10 48550 1706 03741 。/ARXIV。。John Joon Young Chung、Shiqing [4] He 和 Eytan Adar。2021年。《创意支持工具中用户、角色、交互与技术的交汇》。载于*2021*年交互系统设计会议（线上活动，美国）（*DIS '21*）。计算机协会，纽约州纽约市，美国，1817–1833。https: //doi.org/10.1145/3461778.3462050[5] 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绘制故事草图》。载于*2022*年计算机系统中的人为因素会议论文集（美国路易斯安那州新奥尔良）（*CHI '22*）。计算机协会，纽约州纽约市，美国，文章编号209，19页。https: //doi.org/ 10 1145 3491102 3501819 。/。Linda Flower 和 John R. Hayes。1981年。《写[6] 作认知过程理论》。《大学写作与交流》32卷，第4期（1981年），365–387。http://www.jstor. org/stable/356600[7] Deep Ganguli、Danny Hernandez、Liane Lovitt、Amanda Askell、Yuntao Bai、Anna Chen、Tom Conerly、Nova Dassarma、Dawn Drain、Nelson Elhage、Sheer El Showk、Stanislav Fort、Zac Hatfield-Dodds、Tom Henighan、Scott Johnston、Andy Jones、Nicholas Joseph、Jackson Kernian、Shauna Kravec、Ben Mann、Neel Nanda、Kamal Ndousse、Catherine Olsson、Daniela Amodei、Tom Brown、Jared Kaplan、Sam McCandlish、Christopher Olah、Dario Amodei 和 Jack Clark。2022年。《大型生成模型中的可预测性与意外性》。载于*2022*年 *ACM* 公平、问责与透明度会议（大韩民国首尔）（*FAccT '22*）。计算机协会，纽约州纽约市，美国，1747–1764。https://doi.org/ 10 1145 3531146 3533229 。/。Katy Gero、Alex Calderwood、Charlotte [8] Li 和 Lydia Chilton。2022年。《基于写作认知过程模型的写作辅助工具设计空间》。载于第一届智能与交互式写作助手研讨会论文集（*In2Writing 2022*年）。计算机语言学协会，爱尔兰都柏林， 11 24 10 18653 1 2022 2 1 2 – 。https://doi.org/ 。/v / .in writing-[9]。Maliheh Ghajargar 、 Jeffrey Bardzell 和 Love Lagerkvist。2022 年。一个红发女子走进酒吧：与人工智能共同创作小说的体验。载于第*25*届国际学术 *Mindtrek* 会议论文集（芬兰坦佩雷）（*Academic Mindtrek '22*）。计算机协会，美国纽约州纽约，230 241 10 1145 3569219 3569418 – 。https://doi.org/ 。 / 。[10] Steven M. Goodman 、 Erin Buehler 、 Patrick Clary 、 Andy Coenen 、 Aaron Donsbach、Tiffanie N. Horne、Michal Lahav、Robert MacDonald、Rain Breaw Michaels、Ajit Narayanan、Mahima Pushkarna、Joel Riley、Alex Santana、Lei Shi、Rachel Sweeney、Phil Weaver、Ann Yuan 和 Meredith Ringel Morris。2022年。LaMPost：为阅读障碍成人设计并评估一款人工智能辅助电子邮件写作原型。载于第*24*届国际 *ACM SIGACCESS* 计算机与无障碍会议论文集（希腊雅典）（*ASSETS '22*）。计算机协会，美国纽约州纽约，文章 24，18 页。https://doi.org/10.1145/ 3517428.3544819[11] Riccardo Guidotti、Anna Monreale、Salvatore Ruggieri、Franco Turini、Fosca Giannotti 和 Dino Pedreschi。2018年。黑箱模型解释方法综述。*ACM* 计算调查51，5，文章 93（2018年8月），42 页。https: //doi.org/ 10 1145 3236009 . / Daphne [12] Ippolito、Ann Yuan、Andy Coenen 和 Sehmon Burnam。2022年。使用人工智能驱动的写作助手进行创意写作：专业作家的观点。*ArXiv* abs/2211.05030（2022年）。[13] Ellen Jiang、Kristen Olson、Edwin Toh、Alejandra Molina、Aaron Donsbach、Michael Terry 和 Carrie J Cai。2022年。PromptMaker：基于提示的大型语言模型原型设计。载于*2022*年计算机系统中的人为因素会议扩展摘要（美国路易斯安那州新奥尔良）（*CHI EA '22*）。计算机协会，美国纽约州纽约，文章 35，8 页。https://doi.org/ 10 1145 3491101 3503564 . / 。 Mina Lee、Megha Srivastava、Amelia [14] Hardy、John Thickstun、Esin Durmus、Ash- win Paranjape、Ines Gerard-Ursin、Xiang Lisa Li、Faisal Ladhak、Frieda Rong、Rose E. Wang、Minae Kwon、Joon Sung Park、Hancheng Cao、Tony Lee、Rishi Bommasani、Michael Bernstein 和 Percy Liang。2022年。评估人类与语言模型的交互。https://doi.org/10.48550/ARXIV.2212.09746[15] Pengfei Liu、Weizhe Yuan、Jinlan Fu、Zhengbao Jiang、Hiroaki Hayashi 和 Graham Neubig。2023年。预训练、提示与预测：自然语言处理中提示方法的系统综述。*ACM* 计算调查55，9，文章 195（2023年1月），35 页。https://doi.org/10.1145/3560815[16] Todor Markov、Chong Zhang、Sandhini Agarwal、Tyna Eloundou、Teddy Lee、Steven Adler、Angela Jiang 和 Lilian Weng。2023年。现实世界中检测不当内容的整体方法。arXiv: 2208 03274 [ ] . cs.CL Donald [17] A. Norman。。日常用品的设计。Basic Books, Inc.，美国。[18] Laura Weidinger、Jonathan Uesato、Maribeth Rauh、Conor Griffin、Po-Sen Huang、John Mellor、Amelia Glaese、Myra Cheng、Borja Balle、Atoosa Kasirzadeh、Court- ney Biles、Sasha Brown、Zac Kenton、Will Hawkins、Tom Stepleton、Abeba Birhane、Lisa Anne Hendricks、Laura Rimell、William Isaac、Julia Haas、Sean Legassick、Geoffrey Irving 和 Iason Gabriel。2022年。语言模型带来的风险分类体系。载于*2022*年 *ACM* 公平性、问责制与透明度会议（大韩民国首尔）（*FAccT '22*）。计算机协会，美国纽约州纽约，214–229。https://doi.org/10.1145/3531146.3533088[19] Tongshuang Wu、Ellen Jiang、Aaron Donsbach、Jeff Gray、Alejandra Molina、Michael Terry 和 Carrie J Cai。2022年。PromptChainer：通过可视化编程串联大型语言模型提示。载于*2022*年计算机系统中的人为因素会议扩展摘要（美国路易斯安那州新奥尔良）（*CHI EA '22*）。计算机协会，美国纽约州纽约，文章 359, 10 10 1145 3491101 3519729 页。https://doi.org/ 。 / 。[20] Tongshuang Wu、Michael Terry 和 Carrie Jun Cai。 2022 。AI 链：通过串联大型语言模型提示，实现透明且可控的人机交互。载于*2022*年计算机系统中的人为因素会议论文集（美国路易斯安那州新奥尔良）（*CHI '22*）。计算机协会，美国纽约州纽约，文章 385，22 页。https://doi.org/10.1145/3491102. 3517582[21] Daijin Yang、Yanpeng Zhou、Zhiyuan Zhang、Toby Jia-Jun Li 和 Ray LC。2022年。人工智能作为积极的写作者：在人机协作小说创作中与生成文本互动的策略。载于*ACM IUI* 研讨会联合论文集，第 10 卷。[22] Ann Yuan、Andy Coenen、Emily Reif 和 Daphne Ippolito。2022年。Wordcraft：使用大型语言模型写故事。载于第*27*届国际智能用户界面会议（芬兰赫尔辛基）（*IUI '22*）。计算机协会，美国纽约州纽约，841–852。https://doi.org/10.1145/3490099.3511105