arXiv:2303.03199v1 [cs.HC] 6 Mar 2023

选择而非控制：用户如何运用故事内提示与故事外提示词生成方式与大语言模型协作写作

Hai Dang
hai.dang@uni-bayreuth.de
拜罗伊特大学
德国巴伐利亚州拜罗伊特

Sven Goller
sven.goller@uni-bayreuth.de
拜罗伊特大学
德国巴伐利亚州拜罗伊特

Florian Lehmann
florian.lehmann@uni-bayreuth.de
拜罗伊特大学
拜罗伊特，巴伐利亚，德国

Daniel Buschek
daniel.buschek@uni-bayreuth.de
拜罗伊特大学
拜罗伊特，巴伐利亚，德国

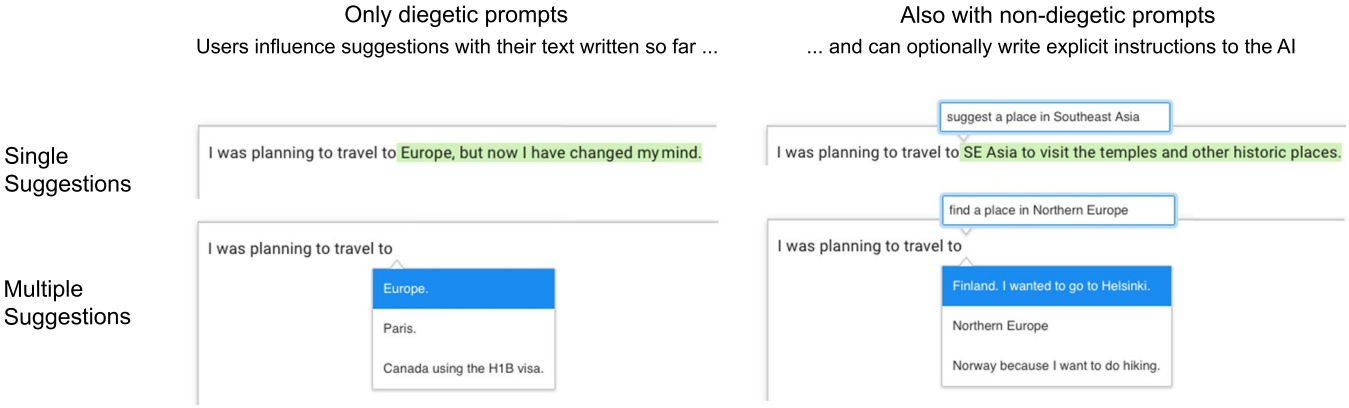

图1：四种用户界面变体概览，展示用户书写的文本（黑色字体，即故事内的提示词）、建议（以绿色突出显示的文本，以及列表中的选项），以及一个弹出文本框，用户可在其中输入指令，将其作为零样本提示发送给系统（即故事外的提示词）。

摘要

我们提出一种关于大语言模型（LLM）提示词的概念性视角，将其分为（1）故事内提示（叙事的一部分，例如“很久很久以前，我看到一只狐狸……”）和（2）故事外提示（外部指令，例如“写一写这只狐狸的冒险经历。”）。基于这一视角，我们研究了*Prolific*上的129名众包工作者如何通过不同的用户界面（1条或3条建议，有无故事外提示；使用*GPT-3*实现）撰写短文本：当界面提供多条建议，并允许使用故事外提示时，参与者更倾向于从多条建议中进行选择，而不是通过故事外提示来控制生成内容。参与者提供故事外提示时，目的是寻求灵感、主题或事实。尤其是，单条建议会同时受到故事内信息和故事外信息的引导。

本研究揭示了以下三点，从而为生成模型的人机交互提供了参考：（1）撰写故事外提示需要付出努力；（2）人们会结合使用故事内提示词生成与故事外提示词生成；（3）他们会利用自己的草稿（即叙事内信息）以及建议时机，有策略地引导大语言模型。

CCS 概念

• **以人为中心的计算** → **人机交互实证研究**；**文本输入**； • **计算方法**。

关键词

大型语言模型、协作式系统、人机协作、以用户为中心的自然语言生成

ACM参考文献格式：
Hai Dang、Sven Goller、Florian Lehmann 和 Daniel Buschek。2023年。选择而非控制：用户如何通过故事内提示和故事外提示与大语言模型写作。载于*CHI '23*：*ACM*计算系统中的人为因素会议，*2023*年*4*月*23*日至*4*月*28*日，德国汉堡。ACM，纽约州纽约，美国，17页。
https://doi.org/10.1145/3544548.3580969

特此授权免费制作本作品全部或部分内容的数字或纸质副本，供个人或课堂使用，但前提是不得为营利或商业利益制作或分发副本，且副本首页须附有本声明及完整引文。对于本作品中由非作者拥有版权的组成部分，必须遵守相关版权要求。注明出处后进行摘要是允许的。除上述方式外，如需复制、再版、发布到服务器或重新分发至列表，必须事先获得具体许可和/或支付费用。请发送邮件至permissions@acm.org 申请许可。*CHI '23*，*2023*年*4*月*23*日至*4*月*28*日，德国汉堡© 2023年版权归所有者/作者所有。出版权已授权给 ACM。ACM ISBN 978-1-4503-9421-5/23/04...$15.00 https://doi.org/10.1145/3544548.3580969

1 引言

在协作写作时，人们通过草稿中的文字以及草稿之外的交流来相互协调、激发灵感。本文探讨了人类与人工智能协同写作中的相关机制。

作为生成文本依据而输入给大语言模型（LLM）的文本，称为“提示词”。在这样的文本提示中提供若干输入和输出示例，有助于模型完成任务 [5, 50]。这称为少样本学习。例如，可以通过提供几个英语句子及其对应的法语译文，再附上需要翻译的英语句子，提示大语言模型将英语翻译成法语。大语言模型随后通过补全文本，（理想情况下）就能翻译该句子。这赋予了用户控制权：用户可以定义任务，并按需将其委托给大语言模型。更进一步，零样本学习通过一条指令来提示大语言模型，而不提供示例（例如，将“天气真好”翻译成法语）。这是一项更难的任务，但从人机交互（HCI）的角度来看，它让用户在指示人工智能系统时，不必再费心思考具体示例。

我们引入故事内提示词生成和故事外提示词生成这两个术语[1]，以构建一个全新的视角，探讨用户如何在写作过程中影响大语言模型。故事内提示词是用户叙事的一部分。例如，当用户撰写一篇关于东南亚度假的文章时，迄今为止写成的故事构成了故事内提示词。相比之下，故事外提示词是给大语言模型的明确指令（例如“建议在新加坡可以开展的活动”）。关键在于，这条指令并不属于最终生成的文档（例如旅行博客）；它仅用于引导大语言模型进行文本生成。

从技术上讲，对于大语言模型而言，故事内提示词和故事外提示词之间可能并无区别——这两种类型都会以文本输入字符串的形式被模型接收。然而，从人机交互的角度来看，这一区分使我们能够识别用户使用大语言模型进行写作时在认知和交互方面的模式。基于这一新区别，本文探讨以下研究问题：用户如何使用故事内提示词生成和故事外提示词生成与大语言模型进行写作？

具体而言，我们提出并比较了四种用户界面变体（图1），让人们能够使用这些类型的提示词进行写作，此外还设置了一个不提供建议的基线用户界面。我们在 Prolific 上对129名众包工作者开展了远程研究，每位参与者都写了五个故事。我们研究了两个自变量对用户写作行为的影响，分别是具有两个水平的指令（i_{no}、 i_{yes}）和具有三个水平的建议数量（基线： s_0； s_1、 s_3）。

总体而言，用户更喜欢从多条建议中进行选择，而不是通过叙事外提示来控制大语言模型。他们会使用叙事外提示，向大语言模型索取灵感、主题或事实。叙事外提示会增加使用者的负担，因为他们需要学习如何组织这类提示词，还要在叙事内写作和叙事外写作之间切换。用户也更喜欢能够看到多条建议的用户界面，而不是只能看到单条建议；不过，允许他们提供叙事外提示，可以提高仅提供单条建议的用户界面的接受率，从而缩小两种界面接受率之间的差距。此外，单条建议通常在句子靠后的位置触发，在过渡词处触发以及用于开启句子的频率也较低。结合用户的评论，这表明写作者会将叙事内信息视为引导大语言模型的依据。我们讨论了这些发现对大语言模型和交互设计的启示。

[1]https://www.merriam-webster.com/dictionary/diegetic, 最后访问于*2023*年*3*月*7*日

我们为提示词生成提出了一种新的概念视角，区分用户影响大语言模型的叙事内方式和叙事外方式；此外，我们还设计了一种新的用户界面，将文本续写建议与零样本提示输入结合起来。

2 相关工作

我们的工作与自然语言处理（NLP）中的提示词生成，以及人机交互（HCI）中的写作界面相关。此外，我们通过将所提出的故事内提示和非故事内提示概念置于现有的写作和提示词生成用户界面中进行讨论。

2.1 大语言模型中的提示词生成

语言模型经过训练，能够根据文本中已有的词语预测下一个词。基于深度学习的大语言模型的一项主要优势在于，它们无需针对特定任务进行专门训练，就能完成多种自然语言处理任务。这可以通过用自然语言编写的文本提示来实现 [5]。Zhao 等人 [50] 表明，提供少量输入和输出示例有助于引导模型。然而，优化提示词并非易事，需要丰富的经验 [28]。

2.1.1 提示工程。 提示工程中的相关工作提出了多种改进提示词的方法：例如，对提示词进行释义可以带来更好的模型输出 [20, 23, 27, 49]。另一种方法是构建提示词模板，以提高知识探测的准确性 [33], 、翻译任务的准确性 [5],，或文本分类任务的准确性 [39]。然而，在提示工程过程中构建的优化提示词通常并非供人类使用；相反，它们旨在让大语言模型以最有效的方式完成任务 [28]。与此不同的是，在本研究中，我们探究非专业用户在使用大语言模型进行写作时，如何编写和使用（零样本）提示词。

2.1.2 提示词生成界面。 已有多种交互系统被提出，以帮助用户更有效地使用提示词：例如，Wu 等人开发的 *AI Chains* [47] 允许用户组合多个提示原语及其输出，形成能够解决复杂语言处理任务的提示词链。在另一项研究中，他们引入了一个用于以可视化方式编程这些链的界面 [46]。类似地，*PromptMaker* [22] 允许用户利用语言提示词来原型化新的人工智能功能。Stro- belt 等人 [42] 开发了一个提示编程环境，使用户能够试验提示词的不同变体，并将提示词的表现可视化。Swanson 等人的 *Story Centaur* [43]支持用户为创意写作创建少样本示例。按照我们的术语，这些项目将非叙事内 提示词作为交互的主要输出。与此不同的是，我们将非叙事内提示词整合进文本编辑器，并将重点放在写作上。具体而言，我们将用于短语建议的用户界面与用于向大语言模型输入零样本提示的用户界面相结合，并分析用户在写作过程中如何使用这些功能。

2.2 大语言模型的写作界面

这里简要概述涉及大语言模型和文本生成的用户界面的关键设计因素。

2.2.1 建议的范围。 早期工作主要集中于单词建议 [11, 12, 18, 34]。这一范围更看重性能指标，例如减少按键次数，而较长的短语建议 [6, 26, 37] 则更多被视为写作中的新想法 [1]。本文聚焦于此类短语建议。

2.2.2 建议的显示方式。 单条文本建议可以以内嵌方式显示 [4, 8, 17, 48], ，而多条建议则以弹出式列表显示，列表中大约包含三至六条内容 [6, 26]。除此之外，Singh等人[41] 评估了写作者如何使用以图像和声音形式显示的建议。此外，Bhat等人 [4] 使用弹出文本框，在句子中间显示用于插入内容的建议。我们沿用这些设计选择（图1），以内嵌方式显示单条建议，并在弹出式列表中显示多条建议。我们还添加了一个弹出式文本字段，用于输入故事外提示。

2.2.3 隐式与显式触发。 在写作界面中，建议可以通过显式或隐式方式触发。相关工作表明，建议可以在短暂的非活动状态后自动显示 [4, 6] ，也可以由效用函数控制 [24]。此外，近期研究还探索了让用户通过快捷键显式请求建议的设计 [7, 17, 26, 41, 48]。我们也采用了这种设计，通过显式的请求按键，更好地了解用户请求建议的时间和方式。

2.3 现有写作界面中的故事内与戏外提示词生成

在此，我们运用所提出的分析视角，探讨现有系统如何在其写作界面中使用故事内信息和戏外信息。传统上，系统主要使用故事内信息，也就是说，它们仅根据前文来预测文本 [11, 12, 18, 34]。有些系统还加入了其他信息（例如手部姿势、身体动作 [15, 16]）。这些都是向语言模型输入戏外信息的早期示例。在本研究中，我们聚焦于文本形式的故事内信息和戏外信息。

从技术角度来看，对于近期使用大语言模型生成文本建议的系统而言，用户的文本草稿（即故事内文本）与输入语言模型的其他文本（例如给模型的指令，即戏外文本）之间可能并无区别。因此，用户界面未明确提供文本提示词输入功能的系统，实际上默认选择仅将故事内信息作为输入提供给大语言模型 [6, 7, 26, 41]。

相比之下，确实允许用户明确输入提示词的写作界面，通常会混合使用故事内信息和戏外信息。Gero等人 [14] 提出了“火花”（sparks），即由大语言模型生成、旨在激发科学写作新思路的句子。用户为生成这些火花所提供的提示词并不属于最终文本，因此属于戏外信息。类似地，其他系统（例如 *Wordcraft* [48], *LaMPost* [17]）允许用户选取已写文本的一部分，并通过预定义功能对其进行修改（这些功能在内部也会使用提示词生成：例如，“重写所选内容”按钮 + 用于输入提示词的文本框）。在此例中，所选文本属于故事内信息，而提示词模板和用户提供的提示词则属于戏外信息。与此相关，我们将用户撰写的整篇文本草稿视为故事内信息，并允许用户提供戏外的自定义文本提示词，以进一步引导大语言模型。

3 交互概念

这里介绍我们的用户界面和交互概念（另见图1和图8）：它在同一用户界面中紧密整合了故事内和戏外的提示词生成；用户无需将手从键盘上移开，即可使用这两种类型。

3.1 行内（单条）建议（图1上排）

当用户请求一条新建议（TAB）时，建议预览会显示在文本编辑器中当前插入符位置之后。用户可以反复按 TAB 键获取新建议。建议预览会以绿色高亮显示，表示它尚未成为文本的一部分。我们选择这种设计，而不是使用灰显的建议文本（例如 Google 智能撰写 [8] 中采用的方式），因为先导测试表明，灰色文本可能让一些人难以阅读，而且会使可读性更加依赖屏幕亮度设置，而这些设置在我们的线上研究中无法控制。如果用户接受建议（ENTER），预览样式（绿色背景）便会移除，建议文本也会成为文本文档的一部分。用户也可以按 ESC 取消当前建议预览，或者不确认建议而继续输入。通过上述任一方式取消建议后，预览中的建议会从文本编辑器中移除。

3.2 多条建议（图1底行）

我们的系统遵循当前提供多条建议的做法（见第2.2节），并将每条短语建议作为列表中的独立条目显示，共三条。同样，用户可以按TAB 获取建议（也可以重复按下以获取新的建议）。用户可以使用 ↑UP/ ↓DOWN 键浏览列表，并按ENTER键接受一条建议。也可以使用鼠标进行选择。

3.3 叙事外提示的弹出式文本框（图1右栏）

在研究中，我们将叙事外提示描述为“给人工智能的指令”。用户可以像之前一样按TAB 请求建议。此外，他们还可以在插入符位置上方出现的弹出框中输入指令。这样，用户可以选择输入指令，但请求建议时并非必须输入。请求建议时，输入焦点会自动从文本编辑器切换到弹出式文本框，因此用户按下TAB 后即可直接输入指令。用户可以按TAB 或Enter 提交指令。随后，他们可以再次按Enter 接受当前选中的建议。或者，他们也可以修改指令以更新建议。

4 原型实现

这里介绍研究中使用的网页原型的详细信息。屏幕截图请参见图1和附录A。

4.1 Web 系统

该原型使用 ReactJS[2] 和 CKEditor53 实现。客户端发出的每个建议请求都会发送至后端服务器并由其解析；该服务器使用 FastApi[4] 作为轻量级 Web 服务器。服务器将这些请求连同全部已写文本以及给建议模型的指令（见第3节）转发至 OpenAI 的 *text-davinci-edit-001* 模型。我们选择该模型和应用程序接口，是因为据称它经过专门训练，能够接收给定文本以及与该文本相关的（单独）指令。

4.2 语言模型提示词

我们使用了两个默认提示词前缀，从 GPT-3（text-davinci-edit-001）获取句子补全结果：(1) 补全句子。(2) 补全句子并 <用户_指令>。</style}当参与者没有提供明确指令的选项，或用户未提供指令时，系统会自动使用 (1)。如果用户确实编写了指令，则会将其附加到 (2) 后面。例如，如果用户写下指令"建议颜色"，发送给模型的完整指令将是"补全句子并建议颜色"。在预研究期间，我们尝试了其他默认指令，例如"继续"或"继续文本"，以及一些更复杂的指令，但发现它们不太合适（例如，会生成更长的文本或结果不够一致）。我们还进行了后处理步骤，以截取模型输出，只显示生成的续写内容。

4.3 信息框

在用户研究中，我们实现了一个信息框（附录A中的图8），用于说明当前文本编辑器设置的不同功能。具体而言，信息框展示了一张图片，说明用户界面的使用方法，并介绍了可用的操作键。

5 方法

我们采用了以下方法，与人机协作写作领域的相关研究保持一致（例如参见 [6, 26]）。

5.1 问卷

为了解参与者的背景，初始问卷询问了他们的人口统计信息，以及使用写作功能和语言模型的经验。参与者还在每种用户界面变体之后填写了一份问卷（见图3），以提供对各个用户界面的主观反馈。为了进一步收集总体反馈，最终问卷请参与者（可选）开放式地评论系统的改动，以及他们对建议和指令的体验。

5.2 交互日志记录

为详细分析交互行为，我们记录了参与者在写作任务期间的交互事件，即键盘和鼠标事件（见附录 A）。每个事件都包含时间戳、任务编号以及编辑器中的当前文本 。根据事件类型，还会包含相关信息例如文本中的建议触发位置，或给人工智能的指令。

5.3 开放式疑问句的编码

我们采用扎根理论的编码步骤 [10, 29]，分析了最终问卷中的开放式评论，以识别并报告其中逐渐显现的方面：首先，两位研究人员通过归纳法，为20人的数据提出编码。随后，他们对比并归类这些编码，制定了一份通用编码本。接着，他们对最初的20名参与者以及另外32名参与者的数据进行编码，并相互检查对方的编码，同时对编码本进行了少量调整。最后，一位研究人员对剩余数据进行编码，另一位研究人员检查了这些编码。在整个过程中，意见分歧均通过讨论解决。

5.4 对用户撰写文本的评估

我们使用多语言语法和拼写检查工具 LanguageTool5，统计语法和拼写错误的数量。为评估参与者在用户研究期间（参见第6节）对所选写作提示的响应程度，三位研究人员分别审阅了这些故事，并就故事与提示词的关联提供了评论。最后，一位研究人员审阅了所有评论，以确保一致性。

6 用户研究

6.1 研究设计

我们的研究采用被试内设计，包含两个自变量：平行建议的数量，分为两个水平：一条和三条建议（s_1, s_3）；以及是否提供编写指令的机会，分为两个水平（i_{no}, i_{yes}）。由此形成四种包含建议的用户界面变体。此外，我们还加入了一种不含任何建议的基线用户界面（s_0）。这五种用户界面的呈现顺序经过完全平衡。因变量包括交互测量指标和问卷数据。

6.2 参与者

我们先对 6 名参与者进行了预研究，通过直接讨论收集详尽反馈；随后开展了主研究，共有 129 名参与者（M=71，F=57，NB=1）。我们通过 *Prolific*[6] 招募参与者，并筛选出英语书面和口语流利且能够使用带键盘计算机的参与者。参与者报告的年龄范围为 18 至 70 岁，中位年龄为 32 岁。按照平台建议，参与者获得的报酬为每小时 8 英镑。

6.3 流程

研究首先呈现描述页面，其中包含有关收集的数据和《通用数据保护条例》的信息，并遵循研究所规定。参与者表示同意后，便会进入一个页面，查看流程概述及涉及的用户界面变体。随后，参与者按平衡顺序完成五项写作任务，之后填写最终问卷。研究预计耗时45分钟（实际平均耗时为45分41秒）。

[2] https://reactjs.org/, 最后访问于*3月 7, 2023*

[3] https://ckeditor.com, 最后访问于*3月 7, 2023*

[4] https://fastapi.tiangolo.com, 最后访问于*2023*年*3*月*7*日

[5] https://languagetool.org，最后访问于*2023*年*3*月*7, 2023*

[6] https://www.prolific.co/，最后访问于*2023*年*3*月*7*日

6.3.1 主题选择。 对于每项任务，参与者首先选择一个写作主题。与相关工作 [26], 一样，允许重复选择；不过，我们要求他们总共至少选择两个不同的主题。主题顺序也经过随机化，并逐一呈现，以鼓励参与者选择多样化的主题。

Gero等人 [13] 提出了写作支持工具的三项任务，包括故事写作和议论文写作。因此，我们从与Lee等人 [26]相同的来源[8] 中选取了五个创意写作主题[7] 和五个议论文写作主题。

6.3.2 写作任务。我们告知参与者，他们有五分钟时间围绕先前选定的主题写作，并以明确的结尾收笔。说明鼓励参与者尝试所有功能，同时也写出自己的文本。文本编辑器下方显示了计时器。我们说明，计时器并非严格的截止时间，而是提醒参与者何时完成任务。我们将参与者提交故事前的最短用时设为15秒，不过大家实际上仍基本都写了五分钟（见第7.2节）。每项任务结束后，参与者都填写了一份问卷（第5.1节）。

7 结果

此处呈现我们的研究结果。统计检验使用 R [35], , 具体而言，使用线性混合效应模型（LMM；借助 *lme4* [2], *lmerTest* [25] 包）。这些模型通过随机截距，兼顾参与者的个体差异以及其所选主题的类型（创意故事写作、议论写作）。模型的固定效应为指令和数量。此外，我们使用 R 包 *multgee* [44] , 通过广义估计方程（GEE）分析李克特量表结果（即序数数据）。显著性水平报告为 $p < 0.05$。

我们将建议会话 定义为一段连续的建议交互过程，从请求建议开始，到取消或接受建议为止（例如，一个会话可能包括连续按三次“Tab”键来浏览建议）。参与者共触发了 3097 次建议会话。在启用建议的任务中，平均值为 6.47（标准差 4.00）, 与相关工作 [26] 中的结果相当。

7.1 建议接受度

我们将接受率定义为已接受的建议数量除以触发的建议会话数量。我们发现不同用户界面之间存在显著差异（均值：s_1=0.55、s_3=0.74、i_{no}=0 59 i_{yes}=0 69 0 64 .、.）。总体平均接受率为.（标准差 0.29）。带有书面指令的建议请求的平均值与此一致（0.64，标准差：0.33）。我们将接受情况拟合为二项数据，并建立了广义线性混合模型（即针对每条展示的建议，记录其是否被接受），结果汇总于表1（第1行）。图2（左上）展示了描述性数据。总之，与展示三条建议相比，展示一条建议显著降低了建议被接受的可能性；然而，允许用户编写指令显著缩小了这一差距，因为这提高了建议的接受率（s_1 不带指令时的平均接受率为0.45，带指令时为0.65）。

7.2 任务完成时间

我们测量的任务时间从开始任务到提交任务为止（均值：s_0=281、s_1=305、s_3=306、i_{no}=303、i_{yes}=309）。由于写作时间固定，因此我们预计这里不会有太大差异。事实上，针对建议用户界面拟合的线性混合模型并未发现显著影响（表1，第2行）。另一个类似模型比较了建议用户界面与基线（表1，第3行）：我们发现，使用三条建议进行写作所需时间显著长于不使用建议的情况。这与图2（上方中央）的描述性结果一致：参与者遵循任务描述，写作五分钟；使用建议用户界面写作所花的时间略长。

7.3 文本长度

总计，提交的文本包含 87,640 个词语，其中包括已接受建议中的文本。每篇文本的平均词数为 134 词（标准差 54）。我们对词数数据拟合了广义（泊松分布）线性混合模型，以比较四项带有建议的任务（表1，第4行）；还拟合了另一个此类模型，以比较建议用户界面与没有建议的基线（表1，第5行）。结果与图2（右上）中可见的描述性模式一致：总的来说，使用单条建议写作或使用允许提供指令的用户界面写作时，文本显著更短。然而，使用多条建议写作会使文本显著变长。这些差异相当小，约为 6–10 个词语（平均值：s_0=136, s_1=130 s_3=140 i_{no}=138 i_{yes}=131 , , , ）。

7.4 建议请求时刻

我们分析了参与者在什么时刻请求建议。

7.4.1 句后与句中。 我们分析了建议有多少次是在句首开始的（例如“Hello, world! *[Tab键]*”），又有多少次是在句中开始的（例如“Hello world, how *[Tab键]*”）。我们将这些请求作为二项数据拟合了一个广义线性混合模型（即记录每个请求是否发生在新句子的开头），结果汇总于表1（第6行）。图2（左下）展示了描述性数据。总的来说，与显示三条建议相比，显示一条建议显著降低了参与者在新句子开头请求建议的概率（均值：s_1=21.70 %、s_3=31.02 %）。

7.4.2 句子中的词数。 对于在句中提出的建议请求，我们进一步分析了参与者在句子中的多少个词之后提出请求。我们对每篇文本中包含建议请求的句子的平均词数拟合了线性混合模型，结果汇总于表1（第7行）。图2（下方居中）展示了描述性数据。总的来说，与显示三条建议相比，显示一条建议显著增加了提出建议请求前句子中的词数，约增加1.5个词（均值：s_1=10.93、s_3=9.48；即相对增加15.3 %）；而指令似乎没有影响（均值：i_{no}=10.18、i_{yes}=10.34）。请注意，在句中晚1至2个词提出请求是相当显著的，因为这可能会导致用户向系统提供的后续内容约束截然不同（例如“The...”与“The man said...”）。

7.4.3 建议请求中的词语。我们进一步分析了用户在什么词语之后请求建议。具体而言，

[7] https://www.reddit.com/r/WritingPrompts, 最后访问于3月 2023

[8] https://www.nytimes.com/2021/02/01/learning/300-questions-and-images-to-inspire-argument-writing.html,最后访问于2023年3月7日

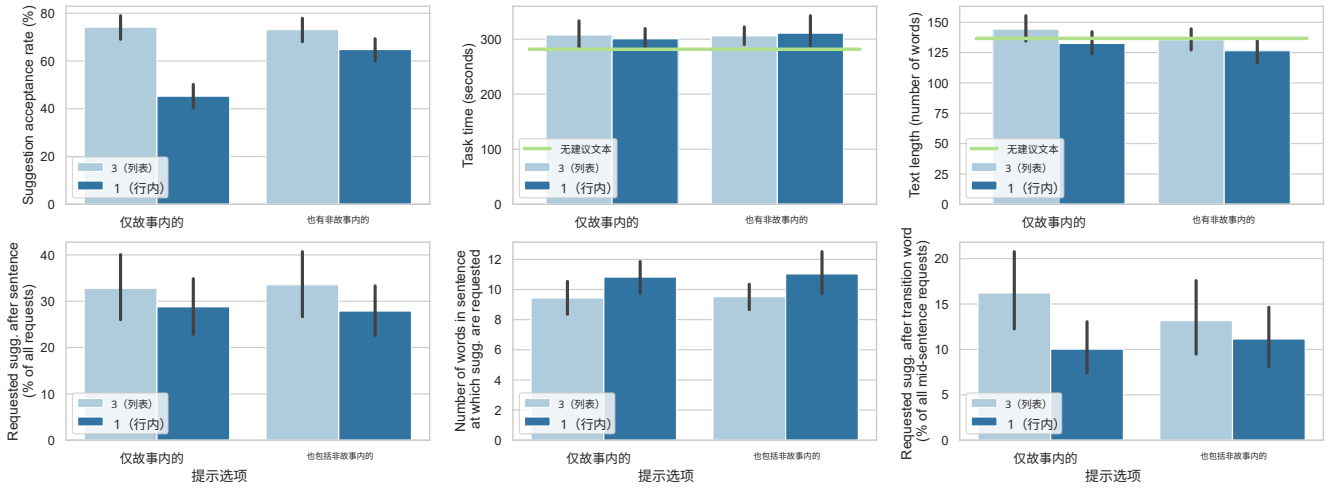

图2：本研究中交互指标概览。总的来说，我们观察到：（1）允许用户选择是否撰写指令（即叙事外提示），会提高单条建议的接受率，但其接受率仍不及多条建议（左上）。（2）单条建议和指令条件下，写作时间变化不大，文本略短（上方中间/右侧）。（3）在句子开头（左下）、句中晚约1.5个词语处（下方中间），以及过渡词之后（右下），用户请求单条建议的频率较低。详情请参阅正文。

我们利用线上英语过渡词列表[9]，将这些"触发词"分为过渡词和其他词。例如，过渡词表示原因（如"because""since"）、对立关系（如"while""despite"）、结果（如"therefore""then"）等。我们在项目代码库中提供了所用词语的完整列表。我们将这些请求拟合为二项数据，并建立了广义线性混合模型（即记录每次请求是否出现在过渡词之后），结果汇总于表1（第8行）。图2（右下）展示了描述性数据。总的来说，与展示三条建议相比，展示一条建议显著降低了用户在过渡词之后请求建议的概率（均值：s_1=11.80 %、s_3=14.03 %）；而指令似乎没有造成显著差异（均值：i_{no}=12.03 %、i_{yes}=13.32 %）。

7.5 对任务和用户界面的感知

我们使用李克特量表条目评估参与者在每项写作任务后的感知（图3）。从描述性结果来看，大多数参与者对建议的评价较为正面，且几乎没有察觉到语法或事实错误。然而，没有任何一种用户界面对所有人而言都显然是"最佳"的：对于各个问题和用户界面变体，意见各不相同，其中也包括对能否撰写指令（即叙事外提示词生成）的有用性的感知。这种差异与参与者所谈及的各种利弊和偏好相符（见第8节）。本文报告广义估计方程分析的结果。由于共有16个问题，我们根据分析所呈现的整体情况进行总结。

7.5.1 对建议数量的感知差异（s_1 与 s_3）。与提供三条建议的列表相比，参与者对仅提供单条建议的评价较差。在若干疑问句上，这一差异具有统计显著性。

[9]例如：https://www.grammarly.com/blog/transition-words-phrases/, https://writingcenter.unc.edu/tips-and-tools/transitions/, 最后访问于2023年3月7日

广义估计方程模型估计，与提供三条建议的列表相比，仅提供单条建议获得较高评分的几率是前者的"x"倍，x如下：单条建议被认为显著更令人分心（x=1.95，p<0.005）、帮助较小（x=0.60，p=0.02）、导致更多手动编辑（x=1.91，p<0.005）、让人感觉控制力较弱（x=0.67，p=0.03），且提供的建议多样性较低（x=0.67，p=0.04）。

7.5.2 对指令（i_{no} 与 i_{yes}）的感知差异。我们发现，人们对指令的感知存在一种权衡：负面来看，允许用户输入给人工智能的指令的用户界面，在"手动编辑建议"的评分上显著更高（x=1.54，p=0.02），且明显更容易分散注意力（x=2.03，p<0.0005）。

正面来看，用户认为给出指令更有助于影响建议文本（x=1.92，p<0.001）。从描述性结果来看，用户也觉得自己更能控制建议文本（见图3中的 Q_{10}），不过这一差异并不显著（x=1.42，p=0.058）。

7.5.3 数量与指令的交互。如前两部分所述，单条建议和提供指令的能力都被认为会显著增加分心程度。然而，指令与数量在分心程度上还存在显著的负向交互效应。与 s_3 相比，s_1 之间分心程度的增加在 i_{yes} 中低于 i_{no}（这也与图3中 Q_2 的情况相符）。这似乎与此前关于接受率的发现（第7.1节）一致：可能是因为通过指令能够找到更多有用的单条建议，从而降低了人们对单条建议和/或指令原本感知到的分心程度。话虽如此，请注意，对于所有建议用户界面，大多数人并不认为它们会造成分心。在分析开放式反馈时（第8节），我们将更详细地讨论分心这一方面。

	章节方面		显著正向预测因素	显著负向预测因素	显著交互作用	用词总结
1	7.1	建议 接受		s_1 (β=-1.26，标准误=0.11，置信区间$_{95\%}$=[-1.48，-1.03]，p<.0001)	数量 * 指令 (β=0.72，标准误=0.17，置信区间$_{95\%}$=[0.39，1.05]，p<.0001)	显示一条建议（而非三条）会降低……的可能性 接受；没有指令时比有指令时更为明显。
2	7.2	任务时间， 比较建议文本。 UIs				在以下对象之间未发现任务完成时间存在显著差异： 四个带有建议的用户界面。
3	7.2	任务时间， 比较建议文本 用户界面与基线 （无建议）	s_3 （β=24.59，标准误=9.5，置信区间$_{95\%}$=[5.96，43.22]，p<.01）			使用三条建议进行写作比不使用建议花费的时间更长 。
4	7.3	文本长度， 比较建议文本。 UIs		s_1 (β=-0.08, 标准误=0.01，CI$_{95\%}$=[-0.10, -0.06]，p<.0001); i_{yes} (β=-0.06，标准误=0.01，置信区间$_{95\%}$=[-0.09，-0.04]，p<.0001)		使用单条建议或 使用允许输入指令的用户界面时，文本略短……
5	7.3	文本长度， 比较建议文本 用户界面与基线 （无建议）	s_3 (β=0.05, 标准误=0.01，置信区间$_{95\%}$=[0.03, 0.07]，p<.0001)	s_1 (β=-0.03, 标准误=0.01，置信区间$_{95\%}$=[-0.05, -0.01]，p<.005); i_{yes} (β=-0.05，标准误=0.01，置信区间$_{95\%}$=[-0.07，-0.03]，p<.0001)		……与基线相比也是如此。然而，使用多条 建议进行写作会使文本稍长。
6	7.4.1	请求 建议后 句子与 句子中		s_1 (β=-0.83，标准误=0.13，置信区间$_{95\%}$=[-1.08，-0.57]，p<.0001); i_{yes} (β=-0.32，标准误=0.14，置信区间$_{95\%}$=[-0.59，-0.05]，p=0.018)		仅显示一条建议（而非三条）降低了 在新句子开头请求建议的可能性。
7	7.4.2	词语数量 请求建议时句子中 的词数	s_1 (β=1.55，标准误=0.78，置信区间$_{95\%}$=[0.01，3.08]，p=0.049)			显示一条建议（而非三条）增加了 请求建议时句子中已有的词语数量。
8	7.4.3	建议请求中的词语类型 建议请求		s_1 (β=-0.37，标准误=0.14，置信区间$_{95\%}$=[-0.65，-0.09]，p=0.010)		显示一条建议（而非三条）降低了……的可能性 在过渡词之后请求建议。

表1：（广义）线性混合模型结果 及显著结果的要点。空白单元格表示没有显著结果。详情请参阅第7.1至7.4节,数据的描述性概览请参阅图2。 **nt**

最后，我们还提出了两个直接关注指令的问题（Q_{15} 和 Q_{16}），因此只有对于带有指令的用户界面才能提出这两个问题（即，图3中 Q_{15} 和 Q_{16}仅有一行故事外的情况）。对于这两个问题，我们没有发现 s_1 和 s_3之间存在显著差异。

7.6 指令的使用与内容

总的来说，参与者使用非叙事式提示选项向系统发送了 397 条指令，平均每人 3.08 条（标准差：3.40）。指令的平均长度为 14.20 个字符（标准差：9.01），即 2.52 个词语（标准差：1.74）。对于允许参与者在请求建议时输入指令文本的任务，参与者输入指令文本的平均比例为 0.19（标准差：0.17）。也就是说，平均每五次建议请求中，就有一次使用了指令。

我们识别出三种主要的指令"风格"：最常见的一种（171次）是使用单个关键词 （或以逗号分隔的关键词列表）。我们还发现了一种祈使句 风格，共出现59次（例如，以"建议""给出""查找""描述"等词语开头撰写提示文本）。有12种情况下，参与者使用疑问句 （例如，以"什么""谁"等疑问词开头，和/或以"？"结尾）。其他情况则包括由多个用于描述某事物的词语（例如"意大利的某处"）组成的指令。从定性角度来看，我们发现了多种方法（表2）。

7.7 文本质量评估

每个词语的拼写和语法错误平均数量为 0.0025，与此前研究报告的数值相当 [26]。约 3.5% 的文本（645 篇中的 23 篇）与选定主题不符。其中大多数（23 篇中的 13 篇）属于"变形者"类别；参与者可能将其误解为对伴侣所期望的一组特质的隐喻。尽管可能存在这种误解，大多数参与者仍认真对待任务，并围绕主题进行了深入反思。

8 开放式反馈

我们按照第5节所述分析了最终反馈，并根据涌现出的方面组织本报告。

8.1 关于建议的评论

大多数人更喜欢多条建议（75 人表示偏好这一方式，23 人偏好单条建议）。主要原因是找到合适建议的机会更多（编码 36 次），以及更多

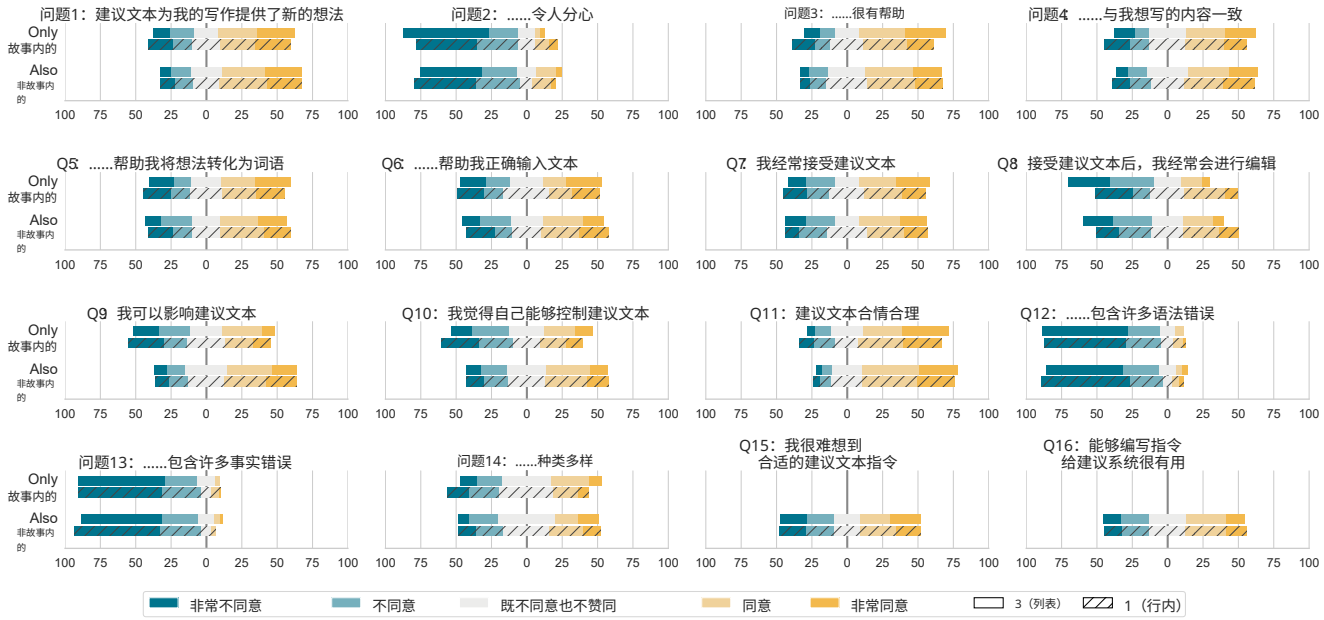

图3：李克特量表结果概览。这些疑问句是在每项写作任务之后提出的。请注意，Q15 和 Q16 与指令（即叙事外提示词）有关，因此仅针对相应的任务提出。

方法	示例
提供主题	“学校”、“书籍”、“退休”、“动物园”、“活动”
提供形容词	“好”、“糟糕透顶”、“坏”、“漫长、太多、疯狂”，“友好”、“有趣”、“害怕”
寻求灵感	“给我讲一个恐怖故事”、“推荐一个地点”、“推荐下一步”、“我们早上做的事情”、“为一名中年男性推荐一项活动”
让想法更加具体	“推荐一些令人恶心的事物”，“爸爸怎么了爸爸”，“推荐一种鸡尾酒”，“推荐一种手枪”
要求提供变体	“另一个短语”，“另一个户外动作”，“推荐一种不同的方法”，“任何东西”
请求写作帮助	“刻板印象的其他说法”，“找一个……的同义词 valuable”，“为年轻人推荐一个词”，“另一个才能的说法”
征求意见/建议	“书籍好吗”，“我接下来该做什么”
检索事实	“最近的星系”、“抗衰老的副作用”、“多瑙河畔的一处地点在多瑙河上”

表2：用户研究期间撰写叙事外提示的不同方法概览。参与者使用单个关键词来建议主题和形容词。多词提示通常采用祈使句风格，表述为疑问句，或表述为没有动词的不完整句子。

灵感（编码 8 次）。正如 P_{38} 所写：“我发现多条建议对用户友好得多，而且由于选项更多，也更能激发灵感。”

偏好单条建议的参与者认为，这种方式更直观（编码 7 次）、使用起来更快（编码 3 次），也不那么容易分散注意力（编码 3 次）：“我非常偏好行内建议，因为它们用起来很直观。”（P_{113}）或者：“我喜欢看到句子在实际位置上的样子，而行内建议让这一点成为可能。”（P_{109}）。还有一些参与者喜欢不用做选择（编码 2 次），但也指出，这可能会导致他们选中不够理想的建议。

14名参与者反思了两者的益处，例如：“一方面，行内建议让我感觉不那么杂乱；如果第一个建议不合适，我只需再次按下*Tab*键。另一方面，同时显示多条建议可能会让我得到一个更好的建议，而我原本可能只会满足于第一个。”（P_{13}）

参与者评论了为何以及如何使用建议，提到了灵感（编码39次）、克服写作障碍（编码3次）或完成句子（编码7次）。例如：“如果建议与我正在写的内容一致，或者我对接下来该说什么有点卡住时，我就会使用这些建议。”（P_{117}）或者：“我倾向于先从自己一个模糊的想法开始，再看看它有什么想法。”（P_{14}）。

十八位参与者明确评论了建议的质量：八位持负面看法（P_{91}：*“[……] 几乎所有内容都得编辑。”*）。九位认为建议时好时坏（P_{130}：“有时 *[建议]* 很有帮助，有时却没有。”）。两位留下了正面评论（P_{27}：“我原本怀疑人工智能是否会与我的想法一致，或提出我确实会使用的措辞，但它大多数时候都做到了，结果让我感到惊喜。”）。P_{109} 进一步指出，当事先知道句子的方向时，“[指令] 帮助人工智能写出更详细、更有趣的句子”。该参与者继续说道，使用“指令中的一两个词来获得更好的句子”后，“[有]时 [确实]有效，但很多时候我最终还是写出了自己的句子，或者大幅修改了建议的句子。”（见图5）。

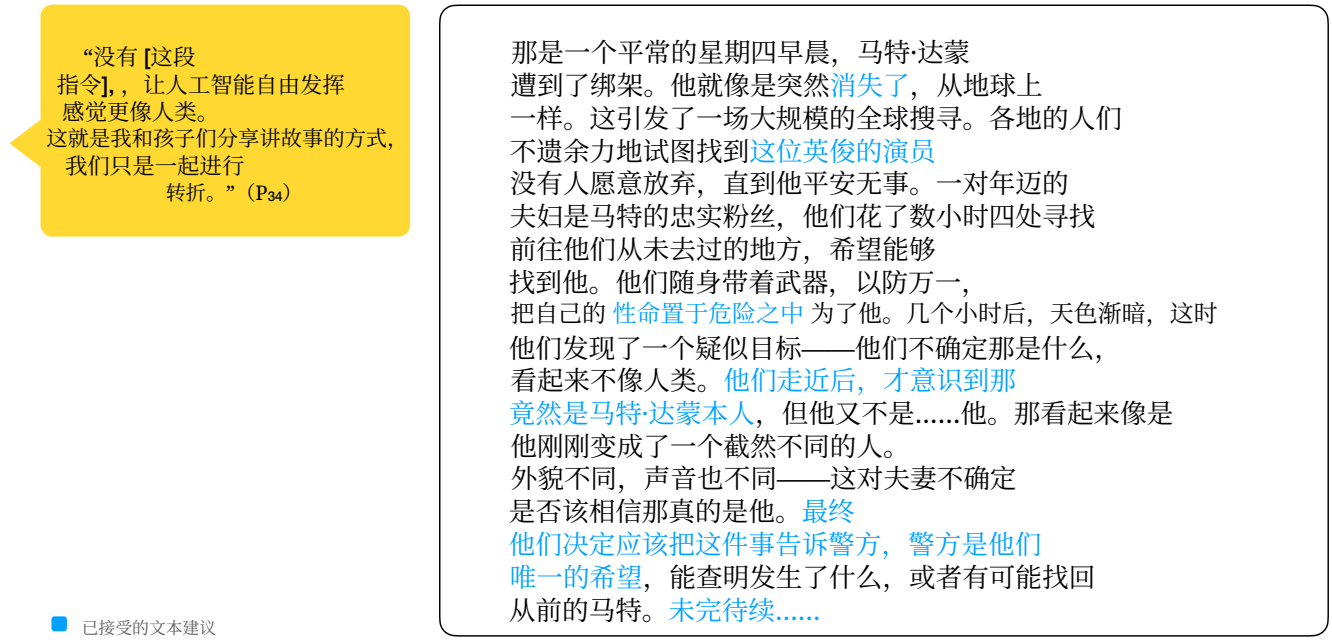

图4： P_{34} **与人工智能轮流写作、讲述马特·达蒙遭绑架故事的文本示例。建议均按原样使用，大多是在句子开头或中间提出的。**

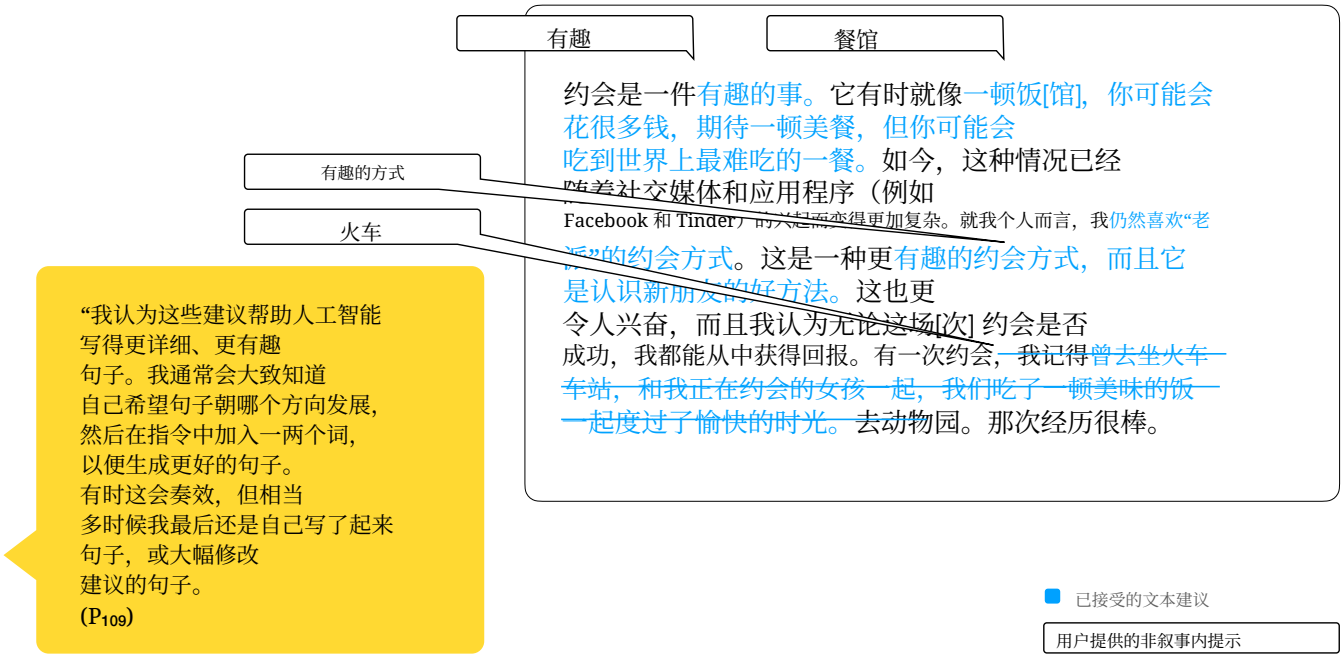

图5： P_{109} **提供非叙事内提示以引导大语言模型的文本示例。对于最后一条指令（“火车”），** P_{109}**随后决定将主题修改为“动物园”。**

8.2 关于指令的意见

参与者对指令的看法不一：21位参与者明确表示，他们更喜欢允许输入指令的用户界面，而24位参与者更喜欢不允许输入指令的用户界面。22位参与者则同时谈到了两者的优缺点。使用指令的主要原因包括：获得更合适的建议（编码12次）（例如 P_{27}：“我发现指令更有帮助，因为我可以在需要时引导人工智能。”）、获取词语灵感（编码29次）（例如 P_{65}：*“[...]* 当我苦于找不到合适的词语时，它给了我灵感。”），以及将构思地点、姓名或同义词等任务交给人工智能。（编码5次）。一位参与者使用人工智能*“[...]* 为我想表达的观点提供正反两方面的建议。”（P_{85}）。

相比之下，有些参与者觉得撰写指令很困难（详情见第8.4节）。六位参与者描述了他们如何采用试错法，以了解怎样才能更好地撰写指令。

还有人报告说，拟定指令会打断写作流程（编码出现3次），从而降低效率，或者不值得花费精力。例如：“这没什么影响，因为我从未觉得有必要给它具体的指令。我觉得它很擅长了解我想要什么样的建议。”(P_{38})。

有人表示，按照指令写作感觉不那么自然（编码出现3次）：“我大部分时候都更喜欢不按照指令写作。我感觉自己更像是与人工智能融为一体，它也更像是与我并肩合作的团队成员，而不是一款软件。我想，这是因为它消除了使用计算机辅助写作的感觉，让我觉得它建议的文字仿佛是我自身的延伸。”(P_{132})。 P_{34} 写道：“我更喜欢不加 *[指令],*，就让人工智能随心发挥。这样感觉更有人情味，就像我和孙辈分享故事时那样，我们轮流讲。”（见图4）。

8.3 控制与影响

十一位参与者评论了控制和影响建议的问题。例如：“我更喜欢*[r]red multiple*，因为——确实——有多条建议可供选择，这让我感觉自己对故事有更好的控制。”(P_{59})。另一位参与者评论道：“我喜欢建议系统，尤其是在我能够提供指导的时候。”(P_{82})。总体而言，参与者提到多条建议和指令有助于增强控制感，这与关于控制和影响的李克特量表结果相一致（图3中的Q_9 和 Q_{10} ）。

此外，参与者还评论了围绕本文如今所称的故事内提示的策略。例如，一些参与者更倾向于通过故事内方法影响建议：“我在提供指令方面并不太成功，很难快速想出建议，因此我转而专注于将主题引向一个能够产生可行建议的地点，而无需交互式输入。”(P_{99})。同样， P_{111} 说道：“很多时候，想出指令和实际写点东西一样困难。似乎直接开始写下我想要的内容，以便将人工智能引导到我希望它前往的方向，会同样容易。”

相比之下，一些人不喜欢故事内提示：“没有指令让人非常恼火，必须调整开头句子，才能让它说出相关内容。这些指令很直观，通常都能正确理解。”(P_{78}，另见图6）。

最后，其他人也注意到故事内提示对自身写作过程的影响：*“[*当我独自写作时，只是漫无边际地写下去；但在与人工智能合作时，我会在脑海中构思该写些什么。我写作是为了能够在创意可能因所提出的建议而走向不同方向的节点请求一条建议。”(P_{99}，另请参见图7）。同样， P_9 写道：*“[*我*]* 注意到，我投入的时间越多，就越倾向于找到一种写作方式，使建议变得有意义，同时也能有趣地添加进去，从而让我的故事更加深入。”

8.4 可学习性

一些参与者（33人）谈到了可学习性和写作指令方面的挑战：“我确实觉得很难想出给 *[*人工智能*]* 的建议。必须先输入句子的开头，然后在一个更小的框里输入我想要的内容，这感觉相当笨拙；考虑到最终生成的内容，这样做也不值得费力。当人工智能理解我的意图并在不需要建议的情况下完成写作时，整个过程流畅得多。”(P_{29}）。 P_6 说：”我几乎因为努力思考该给人工智能什么指令而感到压力，这对我来说真的很难。我很高兴系统提供了这一选项，但我想我并没有充分利用它。” 很贴切的是，25名参与者表示他们没有经常使用指令，例如：*“[.....]* 我不太擅长想出这些指令。”(P_2）。之前的一些评论（第8.3节）也体现了这一点。

8.5 分心

九位参与者明确反思了分心问题。例如P_{20} 写道：“我实际上觉得这些建议相当分散注意力，而且没有帮助。我往往已经知道自己想说什么，因此建议与我的想法契合的可能性相当小。” P_{43}尤其认为指令会分散注意力：“我觉得没有指令的写作能让我更加专注，而且 *[*我*]* 不容易分心，这样我的句子就能更加流畅、自然。如果 *[*我*]* 卡住了，需要帮助，指令就很有用。”。

8.6 对人工智能的感知与期望

这些评论表明了对人工智能角色的两种基本看法：一些人期望系统服务于效率。例如：“我认为这个系统很有价值，但输入指令会让过程变得冗长而繁琐，从而抵消其可能带来的任何好处。相比之下，我更喜欢没有指令的多重建议。”(P_{26})。另请参见上文关于“契合度”的第一条引文（第8.5节）。相比之下，另一些人认为系统应服务于灵感（另请参见第8.2节和第7.6节）。他们向人工智能索要内容建议，或者好奇人工智能会将故事引向哪个方向。这也包括即使不采纳建议，也会从中获得灵感：“我阅读这些建议，有时是为了使用它们，有时只是为了获得关于自己正在写什么的想法”(P_{67})。

9 讨论

9.1 选择与控制

我们的研究为有关通过提示词与用于写作的生成系统进行交互的文献作出了贡献：在简短的创意和议论文写作任务中，参与者总体上更喜欢从呈现给他们的多条文本建议中进行选择，而不是主动编写指令。这一点从多建议用户界面的接受率最高（第7.1节）可以看出；接受率并未因指令而提高，定性反馈也印证了这一点，

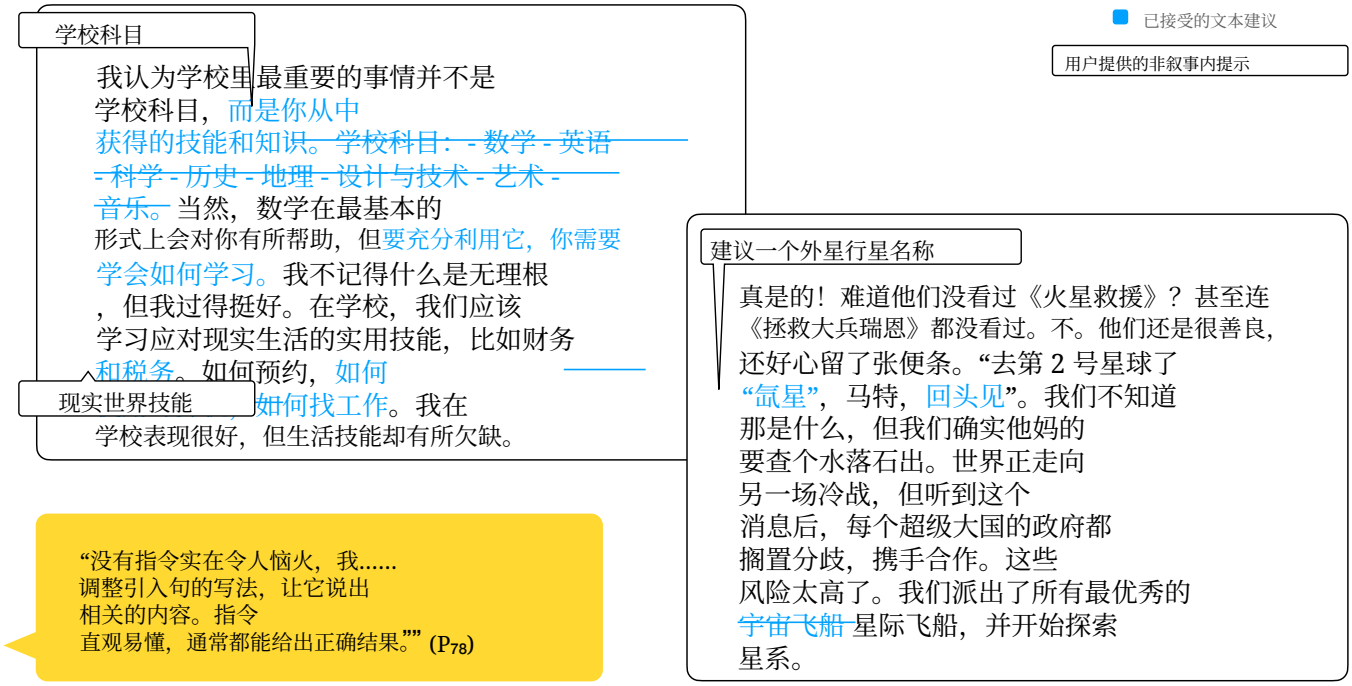

图6： P_{78} **使用故事外提示获取“学校科目”列表的文本示例。已接受的建议以蓝色突出显示。部分已接受的建议后来被删除。**

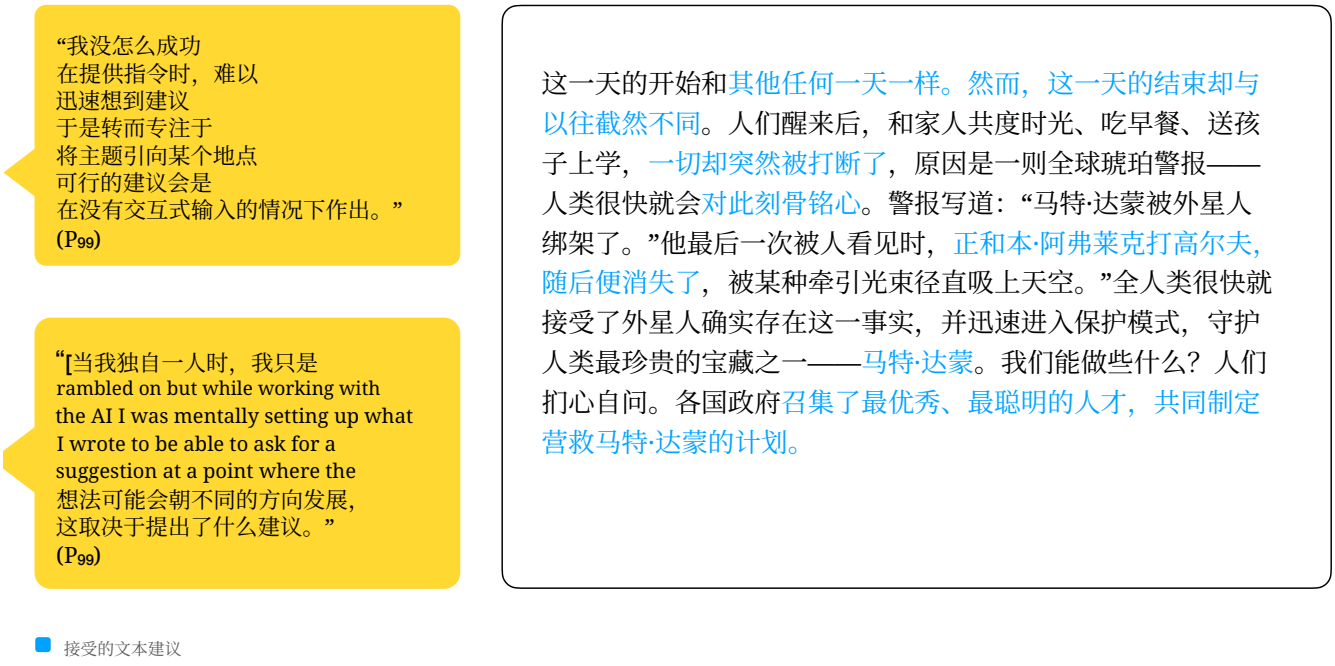

图7： P_{99} **的文本示例，该参与者发现难以提供非叙事内提示词，于是专注于通过叙事内内容引导建议，例如，在“.....送孩子们去上学时”（第3行）之后请求建议。通过在请求建议前以这种方式铺垫句子，该参与者引导大语言模型给出“走向不同方向”的建议（参见第二个黄色框中的评论）。**

其中，明显多数参与者更倾向于多条建议，而对于指令的意见则存在分歧（第8节）。

然而，通过添加故事外提示词生成，在用户界面中提供更多控制选项，部分缓解了缺乏建议选择这一缺点：对于单条建议，指令提高了接受率——尽管这一接受率仍未达到多条建议的水平（第7.1节）。这表明，指令所提供的控制有助于引导单条建议，但其效果不如一开始就提供三条建议供用户选择。

我们讨论几种可能的原因：首先，参与者可能会满足于现状[40]，也就是说，他们会接受一条"足够好"的建议，而不是尝试通过指令对其进行"优化"。建议本身也可能已经足够好，因此不需要指令，这一点得到了一些评论的支持（第8.2节）。其次，一个众所周知的可用性原则是识别优于回忆[31]：与构思一条指令并将其输入相比，用户可能会觉得识别一个已呈现的建议是否合适更加容易。第三，便利性 可能会促使研究参与者在没有指令的情况下接受建议，以便快速完成任务。然而，参与者接受建议的比率与相关工作中的结果相当（该研究提供多条建议，并通过 Tab 键明确提出请求：本研究为74%，[26]中为72%）。对于基于用户指令的建议，我们的比率（64%）高于一项相关研究设计中的比率；在该研究中，用户可以在侧边栏中输入请求（[48]中为17.6%）：这表明，研究设置的潜在影响并不一定会对指令起反作用，或者其影响弱于用户界面设计的影响（例如，侧边栏与集成在文本光标处的设计之间的差异）。此外，时间和文本数据结合评论，进一步支持了参与者认真完成任务这一结论（见第7.7节）。

与此同时，指令确实在特定情境下有用：参与者谈到了它们的益处（第8.2节），每五次建议请求中就有一次使用了指令，并尝试了不同的风格（第7.6节）。综合来看，这些发现促使人机交互领域进一步探索如何通过提示词生成来结合选择与控制。例如，未来的研究可以基于我们的概念视角，构想更多结合故事内提示词生成与故事外提示词生成的用户界面设计，并将我们的数据用作评估基准。

9.2 通过故事内提示词引导建议

我们的研究结果拓展了关于借助人工智能进行写作的文献，揭示了在仅显示单条建议的用户界面中，人们会提供更多叙事内信息，以弥补建议选择不足的问题。这一发现基于首个针对用户在文本何处请求建议所开展的大规模分析：与请求多条建议相比，用户在请求单条建议前的句子中平均多写了约1.5个词。此外，用户较少请求单条建议来开启新句子，或在过渡词之后续写。这或许是因为在这些情况下，收到单条建议的帮助较小：新句子和过渡词都表明叙事的走向可能存在改变的"开放性"。

目前，只有另一项关于触发时刻的小规模分析（N=4，见[7]）。因此，我们鼓励社区在研究具有显式建议触发功能的用户界面时，分析触发时刻。

恰如其分的是，我们最近确实看到，人们对用户显式请求建议的交互设计抱有浓厚兴趣（例如 [7, 26, 41]）。我们的研究通过考察这类设计与建议数量之间的关系，进一步探索了这一设计空间：我们提供了证据，表明人们会考虑何时请求建议；尤其是对于单条建议，他们会在文本中那些预期能为文本续写系统提供更明确指引的位置请求建议。未来的研究可以探讨这一发现是否适用于其他写作情境，以及用户在写作时会在多大程度上主动考虑何时请求建议。根据人们的评论，至少有些人会策略性地思考我们所说的故事内提示词生成（见第8.3节）。

作为一个相关方面，先前的研究聚焦于人们如何回应建议（例如，评估疲劳 [4], 和整合性跃迁 [41]）。与此相辅相成的是，上述结果表明还存在主动的一面：写作者会在看到建议之前就思考这些建议。未来的研究可以对此进行更深入的探讨，尤其是针对用户会明确请求建议的用户界面。

9.3 整合叙事外提示的挑战

我们总结了通过叙事外提示进行交互所面临的两个具体挑战，以指导未来的研究和设计。

9.3.1 叙事外提示会打断写作过程。写作涉及多个认知过程，例如产生想法、将其转化为词语并输入文字 [21]。近期，Bhat等人 [4] 研究了文本建议（不使用叙事外提示）对这一过程的影响。例如，写作者需要评估显示出来的建议。在此，我们的研究进一步揭示了故事内提示与叙事外提示的相对影响：关键在于，从叙事内写作切换到叙事外指令，会迫使写作者从思考自己的叙事或论证转而思考给系统的指令。这一点反映在人们的评论中（第8.2、8.3、8.4节），也体现在有关分心以及难以想到指令的李克特量表结果中（图3中的Q_2 和 Q_{15}）。相比之下，叙事内提示不需要这种思维转换，但仍需要写作者参与评估显示出来的建议 [4, 6]。

9.3.2 叙事外提示词难以撰写。即使完成了这种转变，撰写有效的叙事外提示词仍然很困难，这与相关文献中的发现相呼应 [48]：许多参与者都在这方面遇到困难，并在自我反思中意识到了这一点（第8.2节、第8.3节、第8.4节）。较为积极的一面是，我们在研究中收集到的叙事外提示词展现了用户如何尝试不同的风格。随着使用时间延长，这些风格或许会进一步发展。目前，这些风格都没有超出对人类合著者而言同样有意义的评论范畴。

9.4 对人工智能角色的感知

在此，我们探讨用户如何看待人工智能，并结合 Bhat等人 [4] 提出的写作者与建议交互分析框架，反思三种写作过程，以支持这一讨论：(1)提出新的主题或想法，(2)将抽象思绪或关键词转化为句子，(3) 转录（即输入）词语。

9.4.1 主要角色的两种视角：提议者与转录者。有些人显然将系统视为提供效率的工具

（即转录者），而另一些人则认为它能提供灵感（即提议者）。前者对系统更为挑剔，因为只有当系统速度快、并且能准确预测他们想要什么时，他们才会觉得系统有用。根据定性反馈，我们认为，所选主题以及参与者对该主题的熟悉程度，可能会影响他们的写作心态。在论证性写作中，更广泛地说，当人们对某个主题已有意见时，如果人工智能提出的内容与参与者原本的想法不同，他们就会觉得受到干扰。未来的研究可以进一步考察主题类型和对主题的先验知识如何影响人们对人工智能角色的认知。研究设计在选择写作主题以校准表现或探索指标时，应考虑到这种差异。

9.4.2 故事外提示词反映了用户对人工智能的认知。我们可以进一步探讨故事外提示词的内容如何反映用户对人工智能角色的不同认知：考察写作过程时，我们数据集中的 [4]，个故事外提示词显示，用户会要求人工智能提出富有灵感的想法。有时，用户也只提供不完整的短语或关键词，或询问词语选择，这使人工智能扮演将这些抽象想法转化为完整句子的角色。还有些时候，用户将人工智能视为转录者，以提高输入效率（第8节）。

其他故事外提示词表明，除了这些写作过程之外，还有其他因素会影响用户对人工智能角色的认知：例如，人们会向人工智能征求意见或建议，或要求其查找信息。因此，故事外提示词可能会使用户将人工智能视为写作协作者。

9.5 方法论的局限与反思

参与者使用每项用户界面写作任务的平均时长为十分钟。这与相关工作相当（例如 [26], 11 min [6], 4 min [48]）。未来研究应考察长期使用，尤其是观察随着写作者逐渐熟悉系统，叙事外提示词如何演变。

我们通过应用程序接口使用 GPT-3 对系统进行了原型设计。我们无法直接访问该模型，也不声称已为模型的特定用途找到了“最佳”设置。我们注意到两项局限：有时建议内容重复（例如同一列表中的建议彼此相似），或者重复了指令文本（这似乎没有帮助）。尽管如此，总体而言，参与者对这些建议的评价很高（第 7.5 节）。

模型随时间可能发生的变化不在我们的控制范围内。这限制了此类研究的精确可重复性。我们注意到，学术界部分群体直接访问最先进的大语言模型的机会有限，这一趋势并不容易扭转。积极的一面是，我们的工作表明，可以基于现有模型构建交互式应用，并对其进行详细研究。

我们选择了线上设置，这与近期的相关工作（例如 [6, 26]）一致，以收集许多人的交互日志数据。然而，除了用于完善设计的预研究外，我们无法直接观察参与者，也无法在交互过程中有趣的时刻提出疑问。小样本研究若结合直接观察和出声思维，可以补充我们的工作，例如，更详细地了解触发建议和撰写叙事外提示词时的决策。不过，我们也收到了丰富的定性反馈（第8节）。

指令风格（第7.6节）可能受到提供的示例（附录A中的图8）的影响。我们的预研究表明，这类示例有助于人们开始使用这一新功能。尽管如此，人们还是尝试了示例之外的用法（例如，疑问句、写作帮助、建议等；见表2）。

我们通过弹出框测试了一种整合指令的方式。这个用户界面元素旨在以一种简单的方式，将指令整合到既有的建议列表（或行内建议）设计中。近期的一项相关研究也使用了类似的弹出框（但并非用于指令，而是用于句子中间的建议；参见 [4]）。未来还应探索其他设计。

最后，我们强调开放式写作任务在人机交互研究中的重要性。从历史上看，转录任务一直主导着文本输入研究（参见 [45]）。随着人们对人机协同创作的兴趣日益浓厚，写作工具研究需要新的任务。这些任务不一定要侧重于衡量输入速度，也可以涵盖一系列主题、文本类型及其他方面。从实践角度看，近期研究（例如 [6, 26, 41, 48]）已采用写作者社区的写作任务和定制任务，我们的研究也包括在内。作为一个研究社群，如果要让这类写作任务成为持久且关键的方法学组成部分，我们就应对其进行系统评估和整理。

9.6 超越写作：生成系统中的故事内与故事外交互

我们研究了用于起草文本（即文本到文本）的故事内提示和故事外提示。在这里，我们通过讨论其他交互式生成系统如何运用故事内提示词生成和故事外提示词生成的具体示例，来探讨这一新视角。

- 视觉到文本 Chung 等人 [9] 提出了一种新的故事构思工具，利用视觉草图绘制来引导大语言模型。在这一过程中，草图会被转换为文本提示。这种交互属于故事外交互。
- 文本到视觉 近期的文本到图像模型允许用户根据文本描述生成图像 [30, 32, 36]。这些提示词属于故事外提示，因为它们并非视觉内容的一部分。
- 视觉到视觉 Bau 等人 [3] 展示了一个通过“绘制形状”来引导图像模型的示例：用户绘制三角形等简单形状，以象征一座山。随后，图像模型会将这些形状转换为高保真渲染。由于抽象形状通常并非最终结果的一部分，我们将这种交互视为故事外交互。另一方面，Ha 和 Eck [19] 允许用户先绘制图像的一部分（即提供故事内信息），然后让系统继续绘制或完成这幅图。

因此，区分这两种视角，使研究者能够分析用户在与生成式人工智能交互或设计相关系统时的意图和行为。正如下文讨论所示，我们可以利用这一认识，推导出生成模型交互设计方面的启示。

9.7 大语言模型与用户界面的影响

在 Schick 等人最近的研究中， [38], 他们的大语言模型“PEER”经过明确训练，能够遵循与文本修订相关的叙事外提示词。实际上，我们的研究为此提供了人机交互领域的对应研究——探究如何设计用户界面和交互，将大语言模型以此类角色融入写作过程。我们的结果从两个具体方面为人机交互与自然语言处理交叉领域的未来研究提供指引：

首先，基于我们收集的非叙事内提示词，这些大语言模型应接受训练，以理解更广泛的输入。例如，PEER 采用<style id='1'>祈使句风格进行训练，但我们发现<style id='3'>关键词风格更为常见。或者，也可以通过用户界面引导用户使用受支持的风格。

其次，尽管大语言模型正在快速改进，即使是最优秀的模型也无法消除在叙事内写作与非叙事内写作之间切换所产生的认知成本和交互成本。这促使我们进一步研究需要进行此类切换的交互设计，以及使切换更加轻松、高效的潜在途径。

10 结论

我们对这一现象的新认识表明，人们会使用两种提示方式来引导大语言模型生成文本。以往相关工作介绍的系统主要聚焦于叙事外提示，而我们的发现揭示，用户还会思考并塑造自己的文本，通过叙事内信息引导大语言模型。在采用 GPT-3、同时支持这两种方式的用户界面设计中，参与者更倾向于从多条建议中进行选择，而非撰写指令。最后，我们根据研究结果总结出三项关键启示：

第一，向人工智能撰写指令需要付出努力，包括在叙事内写作与非叙事内写作之间切换。第二，人们会结合故事内提示与非叙事内提示，因为两者都有助于生成单条建议。第三，根据我们对用户请求建议时机的分析，以及他们在评论中的自我反思，写作者会策略性地利用自己的草稿（即叙事内信息）和建议时机来引导大语言模型。

我们鼓励未来的研究进一步分析这些提示类型，以开发更好的写作工具，并将其推广至其他领域（例如，与图像生成模型的交互）。为此，我们在此发布原型，以及有关研究和分析的材料：

https://osf.io/qwakj

致谢

感谢 Lukas Mecke 对稿件提出的反馈。本项目由巴伐利亚州科学与艺术部资助，并由巴伐利亚数字化转型研究所（bidt）协调。

参考文献

[1] Kenneth C. Arnold、Krysta Chauncey 和 Krzysztof Z. Gajos。2018年。《预测文本建议中的情感偏差导致写作偏差》。载于第*44*届图形界面会议论文集（多伦多，加拿大）（*GI '18*）。加拿大人机通信协会，滑铁卢，加拿大，42–49。https: //doi.org/ 10.20380 2018 07 。/GI。Douglas Bates、Martin Mächler、Ben Bolker 和 [2] Steve Walker。2015年。使用 lme4 拟合线性混合效应模型。统计软件杂志67，1（ 2015 1 48 10.18637 067 01），–。https://doi.org/。/jss.v .i [3] David Bau、Jun-Yan Zhu、Hendrik Strobelt、Bolei Zhou、Joshua B. Tenen- baum、William T. Freeman 和 Antonio Torralba。2018年。*GAN* 解剖：生成对抗网络的可视化与理解。技术报告 arXiv:1811.10597。arXiv。https://doi.org/10.48550/arXiv.1811.10597 arXiv:1811.10597 [cs]类型：文章。[4] Advait Bhat、Saaket Agashe、Niharika Mohile、Parth Oberoi、Ravi Jangir 和 Anirudha Joshi。2022年。研究作者与建议的交互：一项定性研究，旨在了解作者与契合/不契合的下一短语建议之间的交互。https://doi.org/ 10 48550 2208 00636 。/ARXIV。。Tom Brown、Benjamin Mann、[5] Nick Ryder、Melanie Subbiah、Jared D Kaplan、Pra- fulla Dhariwal、ArvindNeelakantan、PranavShyam、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。https://proceedings. neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf [6] Daniel Buschek、Martin Zürn 和 Malin Eiband。2021年。多项并行短语建议对母语和非母语英语写作者电子邮件输入与写作行为的影响。载于*2021*年*CHI* 计算系统中的人为因素会议论文集（日本横滨）（*CHI '21*）。计算机协会，纽约，纽约州，美国，论文732，13页。https: //doi.org/10.1145/3411764.3445372

[7] Alex Calderwood、Vivian Qiu、Katy Ilonka Gero 和 Lydia B. Chilton。2020年。小说家如何使用生成式语言模型：一项探索性用户研究。。载于*HAI-GEN+ user2agent@ IUI*。

[8] Mia Xu Chen、Benjamin N. Lee、Gagan Bansal、Yuan Cao、Shuyuan Zhang、Justin Lu、Jackie Tsay、Yinan Wang、Andrew M. Dai、Zhifeng Chen、Timothy Sohn 和 Yonghui Wu。2019年。Gmail智能撰写：实时辅助写作。载于第*25*届 *ACM SIGKDD*知识发现与数据挖掘国际会议论文集 （美国阿拉斯加州安克雷奇）（*KDD '19*）。计算机协会，美国纽约州纽约，2287–2295。https://doi.org/10.1145/ 3292500.3330723[9] 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 *CHI*计算系统中的人为因素会议论文集（*CHI '22*）。计算机协会，美国纽约州纽约。https://doi.org/10.1145/3491102.3501819 活动地点：美国路易斯安那州新奥尔良。[10] Juliet M Corbin。1990年。定性研究基础：扎根理论的流程与技巧。Sage。[11] Mark Dunlop 和 John Levine。2012年。针对速度、熟悉度和拼写检查改进的触摸屏键盘多维帕累托优化。载于*SIGCHI* 人机交互系统中的人为因素会议论文集（美国得克萨斯州奥斯汀）（*CHI '12*）。计算机协会，美国纽约州纽约，2669–2678。https://doi.org/10.1145/2207676.2208659[12] Andrew Fowler、Kurt Partridge、Ciprian Chelba、Xiaojun Bi、Tom Ouyang 和 Shumin Zhai。2015年。语言建模及其个性化对触摸屏打字表现的影响。载于第*33*届 *ACM* 人机交互系统年度会议论文集 （大韩民国首尔）（*CHI '15*）。计算机协会，美国纽约州纽约，649–658。https://doi.org/ 10 1145 2702123 2702503 . / . [13] Katy Gero、Alex Calderwood、Charlotte Li 和 Lydia Chilton。2022年。基于写作认知过程模型的写作支持工具设计空间。载于首届智能与交互式写作助手研讨会论文集（*In2Writing 2022*）。计算语言学协会，爱尔兰都柏林，11–24。https://aclanthology.org/2022.in2writing-1.2

[14] Katy Ilonka Gero、Vivian Liu 和 Lydia Chilton。2022年。Sparks：使用语言模型激发科学写作灵感。载于交互系统设计会议（澳大利亚，线上活动）（*DIS '22*）。计算机协会，纽约州纽约，美国，1002–1019。https://doi.org/10.1145/3532106.3533533

[15] Mayank Goel、Leah Findlater 和 Jacob Wobbrock。2012年。WalkType：利用加速度计数据，适应移动触摸屏文本输入中的情境性障碍。载于*SIGCHI*人机交互系统中的人为因素会议论文集（美国得克萨斯州奥斯汀）（*CHI '12*）。计算机协会，美国纽约州纽约，2687–2696。https://doi.org/10.1145/2207676. 2208662[16] Mayank Goel、Alex Jansen、Travis Mandel、Shwetak N. Patel 和 Jacob O. Wob- brock。2013年。ContextType：利用手部姿势信息改进移动触摸屏文本输入。载于*SIGCHI*人机交互系统中的人为因素会议论文集（法国巴黎）（*CHI '13*）。计算机协会，美国纽约州纽约，2795–2798。https://doi.org/10.1145/2470654. 2481386[17] Steven Goodman、Erin Buehler、Patrick Clary、Andy Coenen、Aaron Michael Donsbach、Tiffanie Horne、Michal Lahav、Bob MacDonald、Rain Breaw Michaels、Ajit Narayanan、Mahima Pushkarna、Joel Christopher Riley、Alex Santana、Lei Shi、Rachel Sweeney、Phil Weaver、Ann Yuan 和 Meredith Ringel Morris。2022年。LaMPost：面向阅读障碍成年人的人工智能辅助写作电子邮件编辑器原型评估。https://arxiv.org/abs/2207.02308

[18] Mitchell Gordon、Tom Ouyang 和 Shumin Zhai。2016年。WatchWriter：在智能手表微型键盘上进行点按和手势输入，并采用统计解码。载于2016年CHI计算系统中的人为因素会议论文集（美国加利福尼亚州圣何塞）（CHI '16）。计算机协会，美国纽约州纽约，3817 3821 10 1145 2858036 2858242 – 。[19] https://doi.org/ 。/ 。David Ha 和 Douglas Eck。2017。。草图绘画的神经表征。http://arxiv.org/abs/1704.03477 arXiv:1704.03477 [cs, stat]。[20] Adi Haviv、Jonathan Berant 和 Amir Globerson。2021年。BERTese：学习与BERT对话。载于第16届计算语言学协会欧洲分会会议论文集：主会场。计算语言学协会，线上，3618–3623。https://doi.org/10.18653/v1/2021. eacl-main.316[21] John R. Hayes。2012年。写作的建模与重塑。《书面交流》29, 3（2012年7月），369–388。https://doi.org/10.1177/0741088312451260[22] Ellen Jiang、Kristen Olson、Edwin Toh、Alejandra Molina、Aaron Donsbach、Michael Terry 和 Carrie J. Cai。2022年。PromptMaker：使用大语言模型进行基于提示的原型设计。载于CHI计算系统中的人为因素会议扩展摘要。1–8。[23] Zhengbao Jiang、Frank F. Xu、Jun Araki 和 Graham Neubig。2020年。我们如何知道语言模型掌握了什么？《计算语言学协会会刊》8（2020年12月），423–438。https://doi.org/10.1162/tacl_a 00324_[24] Anjuli Kannan、Karol Kurach、Sujith Ravi、Tobias Kaufmann、Andrew Tomkins、Balint Miklos、Greg Corrado、Laszlo Lukacs、Marina Ganea、Peter Young 和 Vivek Ramavajjala。2016年。智能回复：电子邮件自动回复建议。载于第22届ACM SIGKDD知识发现与数据挖掘国际会议论文集（美国加利福尼亚州旧金山）（KDD '16）。计算机协会，美国纽约州纽约，955–964。https://doi.org/10 1145 2939672 2939801 。/ 。Alexandra Kuznetsova、Per B. Brockhoff 和 Rune H. B. Christensen。2017年。lmerTest软件包：线性混合效应模型检验。《统计软件杂志》82, 13 2017 1 26 10 18637 082 13 （），– 。[26] https://doi.org/ 。/jss.v .i Mina Lee、Percy Liang 和 Qian Yang。2022。。CoAuthor：设计人机协作写作数据集，以探索语言模型能力。载于2022年CHI计算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388，19页。https://doi.org/10.1145/3491102.3502030[27] Jiachang Liu、Dinghan Shen、Yizhe Zhang、Bill Dolan、Lawrence Carin 和 Weizhu Chen。2021年。什么样的上下文示例适合GPT-3？http://arxiv.org/abs/2101.06804 arXiv:2101.06804 [cs].[28] Pengfei Liu、Weizhe Yuan、Jinlan Fu、Zhengbao Jiang、Hiroaki Hayashi 和 Graham Neubig。2021年。预训练、提示与预测：自然语言处理提示方法的系统性综述。arXiv:2107.*13586 [cs]*（2021年7月）。http://arxiv.org/abs/2107.13586 arXiv: 2107.13586。[29] Michael J. Muller 和 Sandra Kogan。2012年。人机交互与计算机支持的协同工作中的扎根理论方法。载于《人机交互手册》（第3版）。CRC出版社。页数：21。[30] Alex Nichol、Prafulla Dhariwal、Aditya Ramesh、Pranav Shyam、Pamela Mishkin、Bob McGrew、Ilya Sutskever 和 Mark Chen。2021年。GLIDE：借助文本引导扩散模型实现逼真图像生成与编辑。（2021年）。https://doi.org/10.48550/ARXIV.2112.10741 出版方：arXiv 版本号：3。[31] Jakob Nielsen。1994年。增强可用性启发式原则的解释力。载于SIGCHI人机交互系统中的人为因素会议论文集（美国马萨诸塞州波士顿）（CHI '94）。计算机协会，美国纽约州纽约，152 158 10 1145 191666 191729 – 。[32] https://doi.org/ 。/ 。Or Patashnik、Zongze Wu、Eli Shechtman、Daniel Cohen-Or 和 Dani Lischinski。2021年。StyleCLIP：以文本驱动StyleGAN图像的操控。（2021年）。https: //doi.org/10.48550/ARXIV.2103.17249 出版方：arXiv 版本号：1。[33] Fabio Petroni、Tim Rocktäschel、Patrick Lewis、Anton Bakhtin、Yuxiang Wu、Alexander H. Miller 和 Sebastian Riedel。2019年。作为知识库的语言模型？http://arxiv.org/abs/1909.01066 arXiv:1909.01066 [cs]。[34] Philip Quinn 和 Shumin Zhai。2016 年。文本输入建议交互的成本效益研究。载于 2016 年 CHI 计算系统中的人为因素会议论文集（美国加利福尼亚州圣何塞）（CHI '16）。计算机协会，美国纽约州纽约，83–88。https://doi.org/10.1145/2858036. 2858305[35] R核心团队。2020年。R：一种用于统计计算的语言与环境。R统计计算基金会，奥地利维也纳。https://www.R-project. org[36] Aditya Ramesh、Prafulla Dhariwal、Alex Nichol、Casey Chu 和 Mark Chen。2022年。利用CLIP潜变量进行分层文本条件图像生成。（2022年）。https://doi.org/10.48550/ARXIV.2204.06125 出版方：arXiv 版本号：1。[37] Melissa Roemmele 和 Andrew S. Gordon。2015年。创意辅助：故事写作助手。载于《交互式叙事》（《计算机科学讲义》），Henrik Schoenau-Fog、Luis Emilio Bruni、Sandy Louchart 和 Sarune Baceviciute（编）。施普林格国际出版公司，尚，81–92。https://doi.org/10.1007/978-3- 319-27036-4 8_[38] Timo Schick、Jane Dwivedi-Yu、Zhengbao Jiang、Fabio Petroni、Patrick Lewis、Gautier Izacard、Qingfei You、Christoforos Nalmpantis、Edouard Grave 和 Sebastian Riedel。2022年。PEER：协作式语言模型。https: //doi.org/10 48550 2208 11663 。/ARXIV.。Timo Schick 和 Hinrich Schütze[39] 。。利用完形填空题进行少样本文本分类和自然语言推理。http://arxiv.org/abs/2001. 07676 arXiv:2001.07676 [cs]。[40] 赫伯特·亚历山大·西蒙。1996年。人工科学（第3版）。麻省理工学院出版社，马萨诸塞州剑桥。[41] Nikhil Singh、Guillermo Bernal、Daria Savchenko 和 Elena L. Glassman。2022年。藏起被盗的大象：多模态机器智能在创意写作中的飞跃。*ACM*人机交互汇刊（2022年1月）。https: //doi.org/10.1145/3511599 已接受。[42] Hendrik Strobelt、Albert Webson、Victor Sanh、Benjamin Hoover、Johanna Beyer、Hanspeter Pfister 和 Alexander M. Rush。2022年。利用大语言模型进行临时任务适配的交互式可视化提示工程。http: //arxiv.org/abs/2208.07852 arXiv:2208.07852 [cs]。[43] Ben Swanson、Kory Mathewson、Ben Pietrzak、Sherol Chen 和 Monica Di- nalescu。2021年。Story Centaur：作为创意写作工具的大语言模型少样本学习。载于第*16*届欧洲计算语言学协会分会会议论文集：系统演示。计算语言学协会，线上，244–256。https: //doi.org/ 10 18653 1 2021 29 。/v / .eacl-demos。Anestis Touloumis[44] 。。R 软件包 multgee：用于多项响应的广义估计方程求解器。统计软件杂志64, 8（2015年），1–14。http://www.jstatsoft.org/v64/i08/ [45] Keith Vertanen 和 Per Ola Kristensson。2014年。通过写作任务补充文本输入评估。*ACM*人机交互汇刊21, 2，第8篇文章（2014年2月），33页。https://doi.org/10.1145/2555691[46] 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语言模型提示词。http://arxiv.org/abs/2203.06566 数量：arXiv:2203.06566 arXiv:2203.06566 [cs]。[47] Tongshuang Wu、Michael Terry 和 Carrie Jun Cai。2022年。AI Chains：通过串联大语言模型提示词实现透明且可控的人机交互。载于*2022*年*CHI*计算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，美国，第385篇文章，共22页。https://doi.org/10.1145/3491102. 3517582[48] Ann Yuan、Andy Coenen、Emily Reif 和 Daphne Ippolito。2022年。Wordcraft：使用大语言模型进行故事写作。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，纽约州纽约，美国，841 852 10 1145 3490099 3511105 – 。[49] https://doi.org/ 。Weizhe Yuan、Graham Neubig 和 Pengfei Liu。。BARTScore：将生成文本作为文本生成进行评估。http://arxiv.org/abs/2106 11520 2106 11520 。arXiv: 。[ ] cs 。[50] Tony Z. Zhao、Eric Wallace、Shi Feng、Dan Klein 和 Sameer Singh。2021年。在使用前进行校准：提升语言模型的少样本性能。http://arxiv.org/abs/2102.09690 arXiv:2102.09690 [cs]。

A 附录

在本附录中，我们提供了一张记录的事件表以及其他屏幕截图。

No.	交互事件	描述
1	确认指令事件_ _	用户已确认指令（回车键）
2	取消指令事件_ _	用户已取消指令（按 Esc 键或点击指令框外部）
3	事件：打开指令框_ _ _	用户在“带指令”写作设置中触发了新的建议（按 Tab 键）
4	事件：选择下一条建议_ _ _	用户已选择下一条建议（按下箭头键）
5	事件：选择上一条建议_ _ _	用户已选择上一条建议（向上箭头键）
6	事件_请求_建议	用户已请求新的建议（Tab 键）
7	事件：建议响应_ _	系统已返回建议
8	事件：确认建议_ _	用户已选择并确认一条建议（回车键或鼠标选择）
9	事件：取消建议_ _	用户已取消建议（按 Esc 键或点击建议框外部）
10	事件 _TASK_状态	可以是“任务开始”或“任务完成”
11	事件键按下_	用户按下了一个键，例如“A”或“Tab键”

表3：用户研究中记录的交互事件概览。

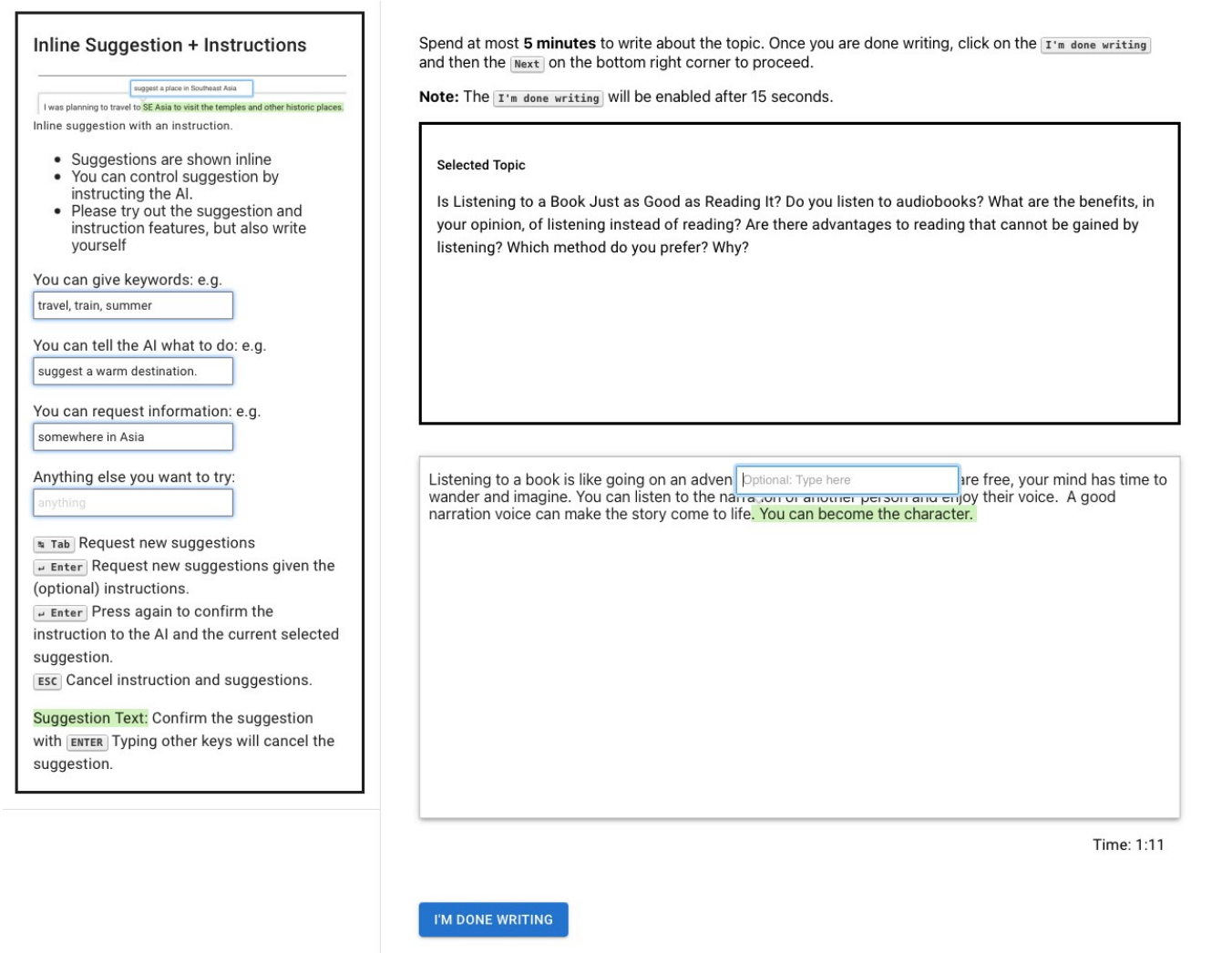

图8：写作界面的屏幕截图。（左侧）信息框描述了当前设置下可用的功能；（中上）选定主题；（中下）文本编辑器，其中显示了当前已写文本和一条行内建议。

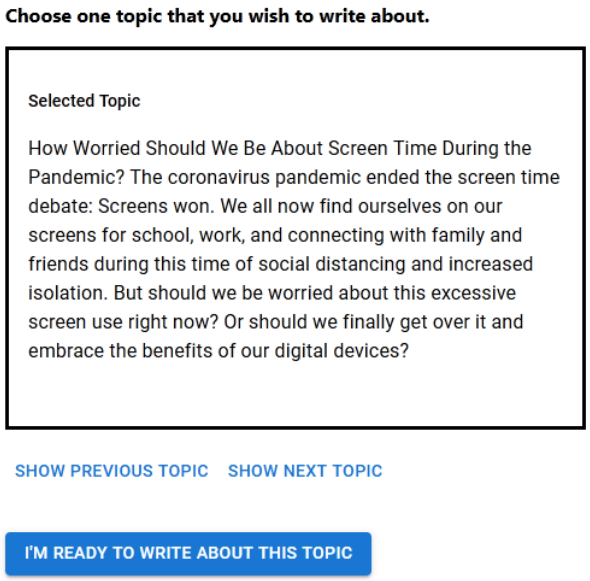

图9：主题选择面板。用户可以浏览主题，并表示自己已准备好围绕所呈现的主题进行写作。