arXiv:2302.00560v1 [cs.HC] 1 Feb 2023

与观点鲜明的语言模型共同写作会影响用户观点

MAURICE JAKESCH, 美国康奈尔大学ADVAIT BHAT, 印度微软研究院DANIEL BUSCHEK, 德国拜罗伊特大学LIOR ZALMANSON, 以色列特拉维夫大学MOR NAAMAN, 美国康奈尔科技学院

如果 GPT-3 这类大型语言模型倾向于生成某种特定观点，它们可能会以未知程度影响人们的意见。本研究探讨语言模型驱动的写作助手（其生成某些意见的频率高于其他意见）是否会影响用户写下的内容以及他们的想法。在一项在线实验中，我们邀请参与者（N=1,506）撰写帖子，讨论社交媒体对社会是否有益。处理组参与者使用了一款语言模型驱动的写作助手，该助手被设置为主张社交媒体对社会有益或有害。随后，参与者完成了一份社交媒体态度调查，独立评审员（N=500）则评估了他们在写作中表达的意见。使用具有观点的语言模型影响了参与者在写作中表达的意见，也改变了他们随后在态度调查中表达的意见。我们探讨了研究结果更广泛的影响，并认为需要更谨慎地监测和设计融入人工智能语言技术中的观点。

ACM 引用格式：
Maurice Jakesch、ADVAIT BHAT、Daniel Buschek、Lior Zalmanson 和 MOR NAAMAN。2023年。与观点鲜明的语言模型共同写作会影响用户观点。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，22页。https://doi.org/10.1145/3544548.3581196

1 引言

像 GPT-3 [21, 97, 103] 这样的大型语言模型正日益融入人类交流。得益于计算机硬件和软件架构的发展， [97], 大型语言模型能够生成类人语言， [56] 它们通过根据前序词语序列反复预测可能出现的下一个词来实现这一点。写作助手等应用 [38], 语法辅助 [66], 和机器翻译 [45] 会将模型的输出融入人们的写作和阅读内容中 [51]。

在日常传播中使用大型语言模型，可能会改变我们形成意见以及相互影响的方式。在传统说服形式中，说服者会精心编写一段有说服力的信息，并将其传递给接收者——无论是面对面交流，还是通过当代技术媒介传播 [94]。近来，用户研究人员和行为经济学家表明，技术性的选择架构（例如选项呈现的顺序）也会影响人们的行为 [42, 72]。随着能够生成类人语言的大型语言模型的出现 [25, 56], , 与技术的互动可能不仅会影响行为，也会影响意见：当语言模型比其他观点更频繁地生成某些观点时，它们可能会说服用户。我们将这种新型的影响范式称为语言模型的潜在说服，如图1所示。

语言模型的潜在说服将“默认选项会影响人们的行为”这一洞见 [42, 72] 拓展到了语言与说服领域。当助推通过让某些选择比其他选择更便利来改变行为时，人工智能语言技术可能会通过让人们更容易表达某些观点、而不是其他观点，来改变意见。这种影响可能是潜在的，且难以准确定位：选择架构是可见的，但内置于语言模型中的观点偏好，对于用户、政策制定者，甚至系统开发者而言，都可能是不透明的。在传统说服中，核心设计者会有意创建信息，以说服特定受众；而语言模型可能会意外地带有倾向，其意见也可能因用户、产品和情境而异。



图1. **传统的技术中介说服（左）与潜在说服通过语言模型实现的潜在说服（右）的比较**。在传统的影响力宣传活动中，中央说服者设计具有影响力的信息或选择架构，并将其分发给接收者。在潜在说服中，语言模型更频繁地产生某些意见，从而影响用户的写作内容，而这些内容又会被他人阅读。

此前关于生成式语言模型风险的研究，主要聚焦于传统的说服场景：人类说服者利用语言模型，自动化并优化广告内容的制作 [39, 61] 或虚假信息 [25, 67, 106]。初步审计还表明，语言模型会再现刻板印象和偏见 [23, 54, 83]，并且比其他文化价值观更支持某些文化价值观 [57]。尽管关于与大型语言模型共同写作的新兴研究表明，模型正日益成为人们写作中的积极合作伙伴 [70, 104, 105],，但人们对语言模型生成的意见如何影响用户的观点仍知之甚少。Arnold 等人的研究[3] 以及 Bhat 等人的研究 [16, 17] 表明，带有偏见的写作助手可能会影响电影评论或餐厅评论，但与大型语言模型共同写作是否会影响用户对公共议题的意见，仍是一个亟待回答的重要开放性问题。

本研究探讨：更频繁地产生某些意见而非其他意见的大型语言模型，是否会影响用户的写作和想法。在一项在线实验中（N=1,506），参与者撰写了一份简短陈述，讨论社交媒体对社会而言是好是坏。处理组参与者看到了由大型语言模型生成的建议文本。模型 GPT-3 [103], 被配置为生成支持“社交媒体对社会有益”的文本，或生成持相反观点的文本。在完成写作任务后，我们通过问卷调查要求参与者评估社交媒体对社会的影响。另有一组人工评审（N=500）对参与者写作中表达的意见进行了评估。

我们的定量分析检验了与具有观点的语言模型的互动是否改变了参与者的写作和问卷调查意见。我们在后续分析中探讨了这种意见转变可能是如何发生的。我们发现，参与者的写作和 他们在问卷调查中对社交媒体的态度，都受到了该模型偏好意见的显著影响。最后，我们讨论研究人员、人工智能从业者和政策制定者如何应对人工智能语言技术可能产生潜在说服的问题。

2 相关工作

我们的研究借鉴了关于社会影响与说服、与写作助手的互动，以及大型语言模型带来的社会风险的既有研究。

2.1 社会影响与说服

社会影响是指个体的思想、感受、态度或行为因与他人互动而发生转变 [92]。社会影响是人类协作不可或缺的一部分，但里程碑式研究表明，它也可能导致不合理或不道德的行为。在个人层面，人们可能会违背自己的判断，顺从多数人的观点 [6] ，甚至服从权威人物，即便这意味着伤害他人 [80]。在社会层面，研究人员发现，社会影响会推动投机市场的发展 [93], 影响投票模式 [69], ，并助长吸烟和肥胖等不健康行为的蔓延 [30, 31]。

随着社交媒体的兴起，线上互动如何影响人们的意见和决策已得到广泛研究。研究表明，多种来源都会影响用户的态度和行为，包括朋友、家人、专家和网红 [48, 78]；后者因能影响大量“粉丝”而被称为影响者 [10]。研究发现，在网络环境中，用户会受到品牌主页、机器人和算法等非人类实体的影响 [41]。研究还评估了个性化推荐、聊天机器人和选择架构等技术产物对用户决策的影响 [15, 35, 50, 72]。

算法实体对人们产生的影响取决于人们如何看待算法，例如，他们是否认为算法建议可信 [50, 76]。随着环境变得更加不确定、决策变得更加困难，算法对个人的影响往往会增强 [20]。随着公众对人工智能发展的认知不断提高，人们可能会将智能 算法视为权威来源 [2, 60, 76]。近期有证据表明，即使在明显出错的简单情况下，人们也可能接受算法建议 [74]。在相关的自动化领域，这种对机器输出的过度依赖被称为自动化偏见 [86, 87, 102]。

2.2 与写作助手的交互

历来，人机交互领域针对文本输入的研究主要关注效率 [68]。典型的文本输入系统会关注词语 [18, 98] 或句子层面的语言上下文 [5, 27]。它们根据底层的似然分布，建议接下来输入一到三个词语 [40, 43, 49, 89]。较新的系统还会提供多条简短的回复建议 [59] 或一条较长的短语建议 [29]。传统上，人们会避免提供更长的建议，因为阅读并选择这些建议所花的时间可能超过手动输入相应文本所需的时间。多项研究表明，自动更正和词语建议等功能可能会对打字表现和用户体验产生负面影响 [11, 26, 37, 85]。

然而，随着规模更大、能力更强的语言模型问世 [21, 97, 103]，，人们对效率以外的设计目标日益关注。研究探讨了界面设计因素，以及如何与写作助手互动，以直接或间接地支持灵感 [17, 70, 95, 105], 、语言能力 [27], 、故事创作 [95, 105], 、文本修订 [36, 107]或创意写作 [33, 46]。在此，语言模型被视为积极的写作伙伴或共同作者 [70, 104, 105]，，而非用于预测或纠正的工具。有证据表明，与提供词语建议的系统相比，提供短语建议的系统 [5] 更容易被用户视为协作者和内容贡献者。

随着写作助手逐渐成为积极参与写作的伙伴，而非单纯的文本输入工具，其“共同创作”对写作过程和产出的影响也会越来越大。Bhat 等人 [17] 讨论了作者如何评估所提供的建议，并将其融入不同的认知写作过程中。Singh等人 [95] 的研究表明，作者会积极努力，或通过“跳跃”，将生成的语言融入自己的写作中。Buschek等人 [27] 提出了九种行为模式，用以表明作者对建议的参与程度各不相同，从忽略建议

到连续串联多条建议。使用建议进行写作，与写出的文本更短、更可预测相关 [4], ；使用语言模型写作时，标准短语的使用也会增加 [17, 27]。此外，建议文本的情感倾向可能会影响所写文本的情感倾向 [3, 52]。

2.3 大型语言模型的社会风险

技术进步催生了新一代语言模型 [21] ，它们生成的语言十分自然，以至于人类几乎无法将其与真实的人类语言区分开来 [56]。得益于计算机硬件和 Transformer 架构的改进 [97], , GPT-3 [23, 90] 等模型因有望影响众多有益的现实应用而备受关注 [21]。然而，也有更为谨慎的声音警告，大型语言模型可能带来伦理与社会危害风险 [100, 101], , 包括歧视和排斥 [23,54, 83] 、虚假信息 [67, 75, 91, 106] ， 以及环境 [96] 和社会经济危害 [14]。

相对而言，较少研究关注生成式语言模型全面部署可能造成的意见、态度和文化普遍变化。已有研究表明，语言模型在其训练所用的语言和语境中（主要是英语或普通话）表现更佳 [23, 91, 103]。小样本审计也显示，GPT-3 等模型中嵌入的价值观，比起其他文化所秉持的价值观，与调查报告中占主导地位的美国价值观更为一致 [57]。Jakesch 等人 [55]的研究指出，开发人工智能系统的人所持的价值观，与使用这些系统的广大人群所持的价值观并不相同。与此相关的人工智能对齐问题——如何构建能够按照运营者的目标和价值观行事的人工智能系统——受到的关注相对更多 [7], , 但研究主要从控制和安全的角度切入。

然而，一个相关主题——社交媒体和推荐系统的政治影响——已受到广泛关注。 [108], 最初，人们对社交媒体的民主潜力充满期待， [62], 但后来显而易见的是，会影响公众舆论的技术将受到强大的政治和商业利益的左右。 [22]算法不再只是技术平台，而是成为公共生活的构成要素， [47] 可能改变政治格局 [1]。研究发现，即使算法并非旨在改变意见，也可能通过强化分歧性意见，加剧政治两极分化 [9, 24, 32]。

3 方法

为了探究与带有倾向性的语言模型互动是否会改变人们的写作并影响他们的观点，我们开展了一项在线实验，邀请参与者（N=1,506）在模拟的在线讨论中使用写作助手回应一条社交媒体帖子。该写作助手所使用的语言模型经过配置，会生成支持某一方论点的文本。我们将参与者的文章和意见与对照组进行比较；对照组在没有写作辅助的情况下撰写社交媒体帖子。

3.1 实验设计

为了在可能真实且相关的情境中研究模型观点与参与者意见之间的互动，我们构建了这样一个场景：在 Reddit 等社交媒体平台上进行带有明确观点的讨论。这类讨论拥有大量读者 [79],涉及政治争议，并且是写作助手和语言模型可能的应用场景。我们检索了 ProCon.org[1],这一针对争议性议题开展研究的在线资源，以确定一个讨论主题。我们选择“社交媒体对社会有益吗？”作为讨论主题。之所以选择这一主题，是因为它是一个

[1]https://www.procon.org/

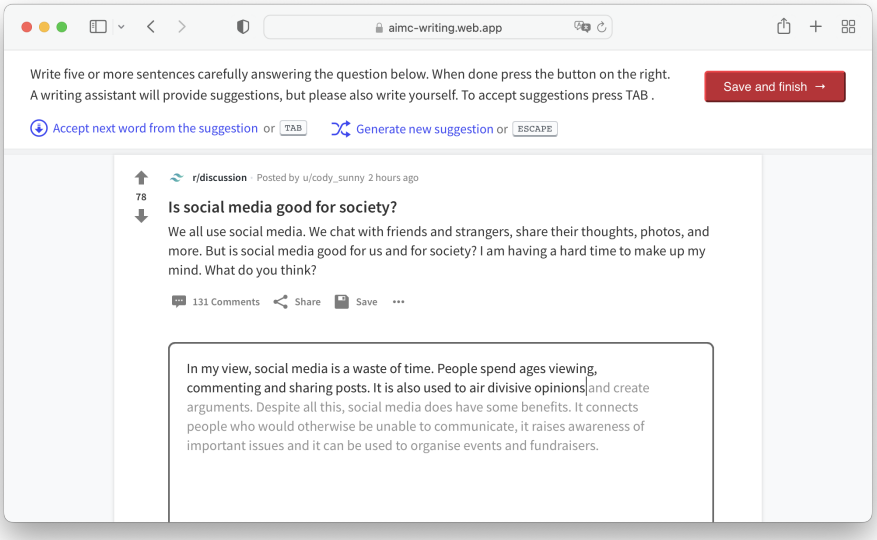

图 2. **写作任务的截图**。页面顶部显示任务，下面是写作助手的使用说明。接着，参与者阅读一则 Reddit 风格的讨论帖，并被要求作出回复。写作助手会显示延续参与者文本的写作建议（以灰色显示）。截图中的参与者写了一段批评社交媒体的论点，但模型被设置为论证社交媒体对社会有益。

容易参与的讨论主题，具有政治相关性，但又不至于被认为争议性过强，从而使根深蒂固的观点限制建设性辩论。

为了开展实验，我们创建了一个定制的实验平台，结合了仿社交媒体讨论页面、富文本编辑器和写作助手。该助手由语言生成服务器驱动，并配备了全面的日志记录工具。为了提供逼真的社交媒体仿制页面，我们复制了 Reddit 讨论页面的设计，并根据 ProCon.org 的讨论主题拟定了一个问题。图 2 展示了实验截图。我们要求参与者至少写五个句子，表达他们对社交媒体对社会影响的看法。我们将参与者随机分配到三个不同的实验组：

(1) 对照组： 参与者在没有写作助手的情况下撰写答案。(2) 技术乐观主义语言模型处理组： 参与者看到的是由配置为论证社交媒体有益于社会的语言模型提供的建议。(3) 技术悲观主义语言模型处理组： 参与者收到的是由配置为论证社交媒体有害于社会的语言模型提供的建议。

3.2 构建写作助手

与谷歌的 *Smart Compose* [29] 和微软 Outlook 中的预测文本类似，实验组中的写作助手会为参与者输入的文本建议可能的续写内容（有时称为“补全”）。我们将这些建议集成到 Quill.js[2] 富文本编辑器的定制版本中。每当参与者暂停写作一段时间（750 毫秒），客户端就会向服务器发送生成请求。包括往返传输和生成时间

[2]https://quilljs .com/

在参与者暂停写作后约 1.5 秒，建议便会显示在他们的屏幕上。

当编辑器客户端收到来自服务器的文本建议时，它会以随机延迟逐字显示建议，这些延迟经过校准，旨在模拟共同写作过程（参见 [71]）。到达建议句子的末尾后，编辑器会暂停，并向服务器请求扩展生成，直到至少建议出两个句子。参与者可以按 Tab键或点击界面上的接受按钮来接受每个建议词。此外，他们还可以通过按下按钮或按键重置生成，以请求新的建议。

我们将所需的云函数、文件和交互日志托管在谷歌的 Google Firebase 平台上。

3.3 配置具有鲜明观点的语言模型

在本研究中，我们尝试使用强烈 倾向于某种观点而非另一种观点的语言模型。我们选择了力度较大的操纵方式，因为我们想要探索语言模型影响 用户意见的潜力，并了解它们是否能被用于或滥用，以改变人们的观点 [8]。

我们使用 GPT-3 [23] 和人工设计的提示，实时生成实验所需的文本建议。具体而言，我们使用了 OpenAI 功能最强大的 175B 参数模型（“text-davinci-002”）。我们采用温度采样，这是一种受统计热力学启发、从一组可能性较高的下一个词元中选取特定下一个词元的方法。我们将采样温度（随机性参数）设为 0.85，以生成多样且富有创意的建议。我们将频率惩罚和存在惩罚参数设为 1，以降低模型建议变得重复的可能性。我们还设置了 logit偏置参数，降低相应词元被选中的概率，从而防止模型生成新行、占位符和列表。

我们评估了多种创建带有鲜明立场的模型的技术，也就是说，这种模型在生成建议时很可能支持辩论中的某一方。我们采用了提示设计 [73], ，这是一种引导冻结语言模型执行特定任务的技术。我们没有更新底层模型的权重，而是将工程化提示与输入文本拼接起来，以提高模型生成特定意见的可能性。具体而言，在生成续写建议时，我们会在参与者的书面文本前插入前缀“社交媒体对社会有益吗？请解释社交媒体为何对社会有益/有害：”。参与者在编辑器界面中看不到工程化提示；它会在生成前由后台插入，并在向参与者展示生成文本前移除。

初步实验和验证表明，该提示能让生成文本表达预期的意见；但当参与者在写作中强烈主张另一种意见时，模型生成的续写就会顺着他们的意见。因此，除了前缀提示，我们还设计了一种中缀提示，在参与者的写作中插入提示，以强化预期的意见。我们在参与者已写内容的最后一句之前插入了片段（“用一句话续写文章，解释社交媒体为何有益/有害：”）。即使参与者在此前的写作中表达了不同意见，这条额外提示也能引导模型的续写朝目标意见倾斜。模型意见配置的验证见第4.5节。我们还尝试了微调 [53]来引导模型表达意见，但微调后的模型未能稳定地产生预期的意见。

3.4 结果测量指标与协变量

我们收集了不同类型的结果测量指标，以探究参与者的观点与模型观点之间的相互作用：

帖子中表达的观点：为了评估表达的观点，我们将参与者的书面文本拆分为句子，并请众包工作者评估每个句子中表达的观点。每

位众包工作者评估了 25 个句子，判断每个句子是否主张社交媒体对社会有益、有害，或利弊兼有。对于主张两者皆非或与此无关的句子，我们还提供了第四个标签。例如，“社交媒体还会助长网络欺凌，导致自杀事件增加”（*P#421*）被标记为主张社交媒体对社会有害，而“社交媒体也有助于营造社区归属感”（*P#1169*）则被标记为社交媒体对社会有益。我们为参与者撰写的每个句子收集了一到两个标签，并为写作助手提供的部分建议收集了标签。对于收集了多个标签的句子，不同评审者给出的标签有 84.1% 的一致率（Cohen’s $\kappa = 0.76$）。

实时写作交互数据：我们全面收集了参与者与模型的建议互动情况，记录精确到每次击键。我们记录了参与者撰写的文本、模型提供的文本，以及参与者采纳了写作助手的哪些建议。我们还测量了他们停下来考虑建议的时间，以及采纳了多少条建议。

意见调查（任务后）：完成写作任务后，参与者填写了一份意见调查。核心问题“总体而言，您认为社交媒体对社会有益吗？”旨在评估参与者态度的变化。为减少需求效应，该问题没有在写作任务结束后立即呈现。我们还提出了一些次要问题，以更详细地了解参与者的观点：“社交媒体如何影响您与朋友和家人的关系？”“社交媒体的使用会导致心理健康问题或成瘾吗？”“社交媒体是否助长了虚假信息和仇恨言论的传播？”“您支持还是反对政府监管社交媒体公司？”这些问题部分改编自 Morning Consult 的全国追踪民调 [34]；答案采用常见的 3 点和 5 点李克特量表。

用户体验调查（任务后）：完成意见调查后，实验组参与者填写了一份问卷调查，反馈他们使用写作助手的体验。他们需要回答“写作助手对你有多大帮助？”，以及“写作助手理解你想表达什么”和“写作助手知识渊博、具备专业知识”这两项陈述是否符合实际。为了了解参与者对写作助手意见的认知，以及这些意见对他们自身观点的影响，我们询问他们是否认同“写作助手提出的建议合理且平衡”，以及“写作助手启发或改变了我的想法和论点”。回答采用 5 点李克特量表，范围从“非常同意”到“非常不同意”。此外，还有一道开放式问题，询问参与者认为写作助手最有帮助或最令人沮丧的地方是什么。

协变量：研究结束时，我们请参与者自行报告年龄、性别、政治倾向和最高学历。我们还构建了一个“模型立场一致性”协变量，用于估计模型支持的意见是否与参与者的意见一致。为避免承诺效应，我们没有询问参与者在写作任务开始前的总体判断，而是在研究结束时询问他们参与讨论前是否认为社交媒体对社会有益。尽管这一指标并不完美，但可作为参与者任务前意见的替代指标。该指标受到所观察到的处理效应影响，其数值相当于标准差的 14%。

3.5 参与者招募

我们为写作任务招募了 1,506 名参与者（排斥后），控制组、技术乐观主义处理组和技术悲观主义处理组分别有 507、508 和 491 人。样本量根据试点研究中任务后问题“总的来说，你认为社交媒体对社会有益吗？”所观察到的效应量计算，检验效能为 80%。样本通过 Prolific [84] 招募。样本包括年满 18 岁的美国参与者（M= 37.7，SD= 14.2）；48.5% 自我认同为女性，48.6% 认同为男性。38 名参与者认同为非二元性别，另有 8 人选择自我描述性别或不透露其性别认同。

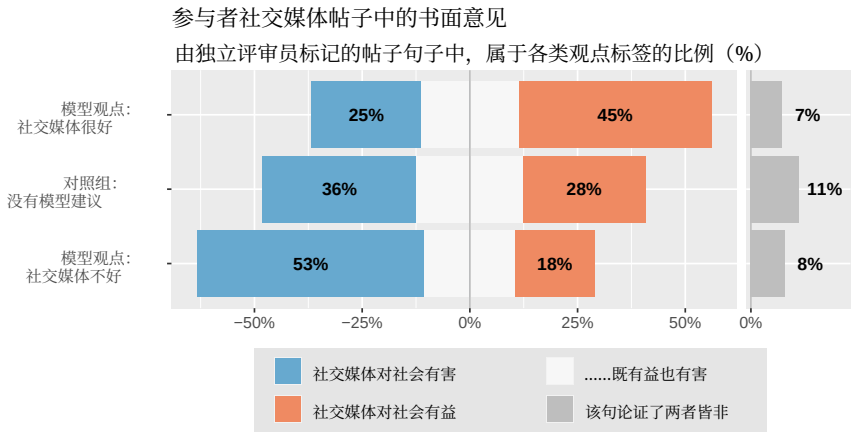

图3. **在社交媒体问题上得到支持性模型协助的参与者，更有可能在自己的帖子中论证社交媒体有益于社会（反之亦然）**。由 Nj=500 名评审员评估了 Np=1,506 名参与者撰写的 Ns=9,223 个句子。纵轴表示参与者撰写社交媒体帖子时，是否借助了支持（上方）或批评（下方）社交媒体的具有观点的语言模型。横轴表示参与者在写作中论证社交媒体不利于社会（蓝色）、有利于社会（橙色）或兼有利弊（白色）的频率。

十分之六的参与者表示具有自由主义倾向；57.1% 至少获得了学士学位。未通过任务前注意力检查的参与者（8%）被排除。获准参与任务的参与者中，有 6% 未完成任务。我们按每小时 15 美元的报酬标准，向参与者支付 1.50 美元；任务平均耗时 5.9 分钟。对于标注任务，我们通过 Prolific 招募了规模相近的样本，共 500 名参与者。实验方案经康奈尔大学机构审查委员会批准。

3.6 数据共享

实验材料、分析代码和收集的数据均可通过开放科学资料库公开获取（https://osf.io/upgqw/）。研究助理对数据进行了筛查，并在发布前删除了可能包含隐私敏感信息的记录。

4 结果

我们首先分析参与者在社交媒体帖子中表达的观点。随后，我们考察参与者是否可能仅仅为了方便而采纳模型的建议，以及模型是否影响了参与者在后续问卷调查中的观点。最后，我们呈现有关参与者对模型观点及其影响的看法的数据。报告的统计结果基于逻辑回归模型。

4.1 与语言模型的互动是否影响了参与者的写作?

图 3 展示了各处理条件下的参与者（纵轴）在写作中认为社交媒体对社会是好是坏的频率（横轴）。对照组（中间一行）参与者撰写的社交媒体帖子略带批判性：他们在 38% 的句子中认为社交媒体对社会有害，在 28% 的句子中认为社交媒体有益。

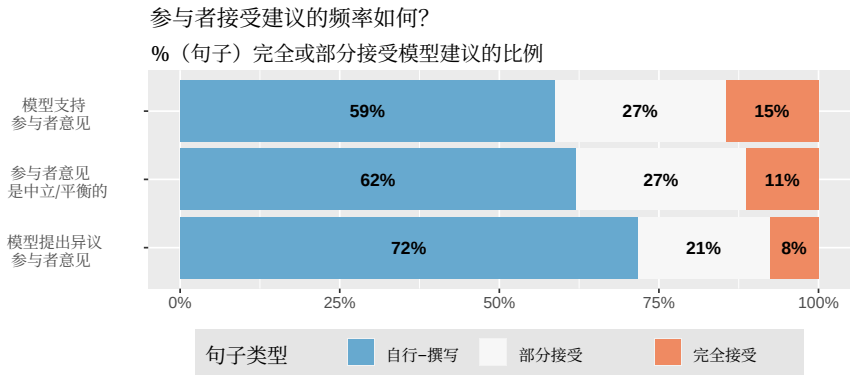

图 4. **如果模型的意见与参与者自身观点一致，他们更可能接受建议**Ns=6，由 Np=1,000 名参与者撰写的 142 个句子。横轴显示参与者自行撰写（蓝色）、与模型共同撰写（白色）或完全接受模型建议（橙色）的句子数量。纵轴根据模型建议是否符合参与者在任务前可能持有的意见，对数据进行了细分。

对照组参与者在约 28% 的句子中认为社交媒体既有益也有害；11% 的句子则认为两者皆非或与此无关。

收到支持社交媒体的语言模型提供建议的参与者（图3上排）主张社交媒体有益的可能性是对照组参与者的 2.04 倍（$p<0.0001$，95% CI [1.83，2.30]）。相反，收到批评社交媒体的语言模型提供建议的参与者（下排）主张社交媒体有害的可能性是对照组参与者的 2.0 倍（$p<0.0001$，95% CI [1.79，2.24]。我们由此得出结论：使用具有观点的语言模型会影响参与者的写作，使他们写出的文本更可能支持模型偏好的观点。

4.2 参与者是否仅仅出于方便而接受模型的建议 ?

参与者可能只是为了方便而接受模型的建议，尽管这些建议与他们原本想表达的内容并不相符。尤其是在在线研究中，付费参与者可能会为了尽快完成任务而倾向于接受建议。

我们的数据表明，在不同实验条件和处理方式下，大多数参与者并未盲目接受模型的建议，而是与模型互动，共同撰写社交媒体帖子。平均而言，参与者自行撰写了 63% 的句子，没有接受模型的建议（见图 4）。约 25% 的参与者句子由参与者和模型共同撰写，通常意味着参与者写下部分文字，并接受模型对该句其余部分的建议。只有 11.5% 的句子完全采用了模型的内容。个人观点可能与模型一致的参与者更有可能接受建议，而持对立观点的参与者接受的建议则更少。约四分之一的参与者没有接受任何模型建议，十分之一的参与者所发帖子的内容有超过 75% 是由模型撰写的。

4.2.1 轻易接受建议是否扩大了观察到的书面意见差异？花在任务上的时间较少的参与者，其写作受到模型的观点影响更大。我们根据参与者撰写帖子的用时，估算他们在多大程度上可能

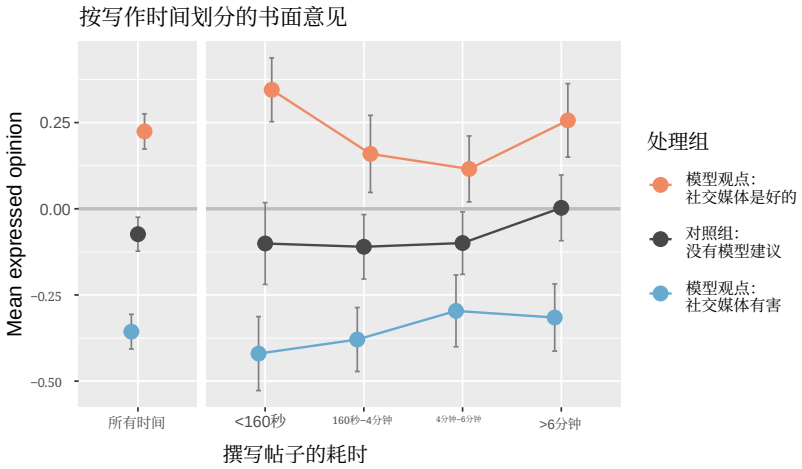

图5。**参与者完成任务越快，其写作中的意见差异就越大。** N=1,506。纵轴显示参与者社交媒体帖子中表达的平均意见，该数值根据汇总的句子标签计算，范围从 -1（“社交媒体对社会有害”）到 1（“社交媒体对社会有益”）。横轴表示参与者花了多少时间撰写帖子。作为参考，左侧面板显示了汇总不同写作时间后得出的表达意见。

未经充分考虑便接受建议。为进行简洁的统计分析，我们将序数意见量表视为区间量表。由于意见量表的各区间大小相近且设有零点，进行连续分析是有意义且合理的 [64]。我们将“社交媒体对社会有害”记为 -1，将“社交媒体对社会有益”记为 1。对于同时支持两种观点或两者皆非的句子，则记为 0。

图5 展示了参与者社交媒体帖子中表达的平均意见，以及该意见如何随实验组别和写作时间而变化。左侧面板显示了参与者在不同时间下表达的意见，供参考；每个实验组与对照组之间的平均意见差约为 0.29（$p<0.001$，95% CI [0.25,0.33],SD=0.58），对应较大的效应量 $d=0.5$）。写作时间较短的参与者（少于 160 秒；右侧面板中最左侧的数据点）受到语言模型意见的影响更大（0.38，$p<0.001$，95% CI [0.31,0.45]）。我们的分析表明，出于便利而采纳建议，促成了书面意见的差异。然而，即使对于花四到六分钟撰写帖子的参与者，我们仍观察到不同实验组之间的意见存在显著差异（0.20，$p<0.001$，95% CI [0.13, 0.27], ，对应的处理效应为 $d=0.34$）。

4.3 语言模型是否影响了态度调查中参与者的观点?

参与者写作中的意见差异，可能是与带有立场的模型互动后，参与者的实际意见发生变化所致。我们评估了与语言模型互动是否影响了参与者在任务后调查中表达的态度。该调查询问参与者是否认为社交媒体对社会有益。参与者回答概况见图 6。

该图展示了每种实验条件下参与者对不同问卷调查选项的选择频率（横轴为答案，纵轴为参与者所处的实验条件）。未与带有立场的模型互动的参与者（图6中间一行）对社交媒体的评价较为平衡：33% 的人回答说社交媒体不

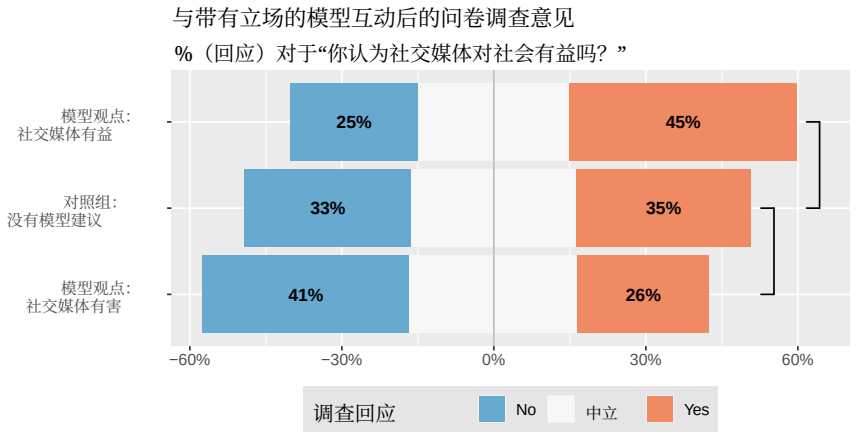

图 6. **在后续问卷调查中，与支持社交媒体的模型互动的参与者更有可能表示社交媒体对社会有益（反之亦然）**。Nr=1,506 份调查回应，来自 Nr=1,506 名参与者。纵轴表示参与者在写作任务期间收到的是支持还是批评社交媒体的模型所提供的建议。横轴显示他们在之后的态度调查中表示社交媒体对社会有益（橙色）或无益（蓝色）的频率。尚未决定的参与者以白色显示。括号表示意见差异显著，显著性水平为 **p<0.005 和 ***p<0.001。

利于社会（中间的蓝色部分）；35% 的人表示社交媒体有利于社会。相比之下，与支持社交媒体的语言模型互动过的参与者（顶行）中，有 45% 的人回答说社交媒体有利于社会。将参与者的答案转换为区间量表后，意见上的这一差异对应的效应量为 d=0.22（p<0.001）。同样，与批评社交媒体的语言模型互动过的参与者（底行）之后更有可能表示社交媒体不利于社会（d=0.19，p<0.005）。

4.3.1 带有立场的模型是否逐渐说服了参与者？虽然我们无法确定说服的机制，但我们的结果进一步揭示了这一过程可能是如何发生的。图7展示了参与者的书面意见在整个写作过程中的演变。在对照组中（黑色），参与者往往先以两句正面陈述开篇，接着写出两句更具批判性的陈述，并以总体上批判性的结论收尾。与对社交媒体持正面评价的模型互动的参与者（橙色），在整段陈述中始终对社交媒体持更正面的评价。与批评社交媒体的模型互动的参与者（蓝色），也写出了更多批评社交媒体的句子，从第一句话开始便是如此。与对照组相似，他们在开头较为正面，在帖子的结尾则更具批判性，这表明写作助手是强化而非取代了他们的叙事。

4.4 参与者是否意识到了模型的观点及其影响？

写作任务结束后，我们询问处理组参与者使用写作助手的体验。我们根据他们的回答，评估他们在多大程度上意识到了模型的观点及其影响。

绝大多数参与者认为语言模型具备专业知识、知识渊博——即使它与他们的个人观点相悖。如图所示8，84% 的参与者表示，当语言模型支持他们的

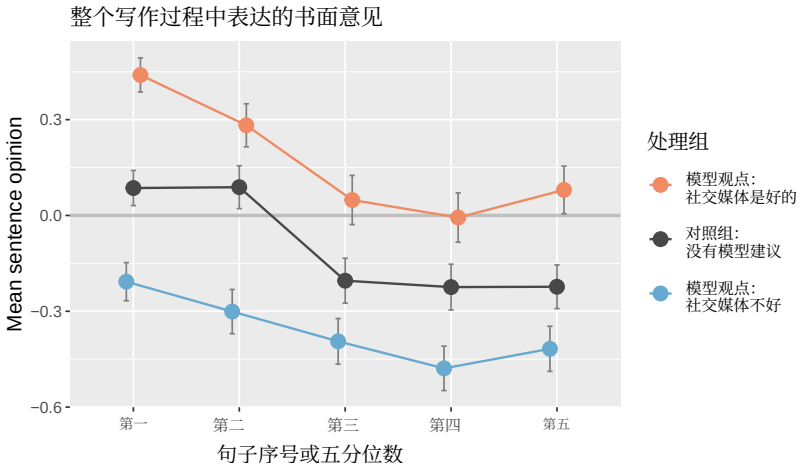

图 7. **在整个写作过程中，模型对参与者写作的影响程度始终相同。**Ns=9，来自 Np=1 名参与者的 223 个句子。纵轴表示参与者句子中表达的意见的平均值。横轴表示该句子在参与者社交媒体帖子中的位置是较早还是较晚。由于大多数参与者按要求写了五个句子，五分位数大致对应句子序号。

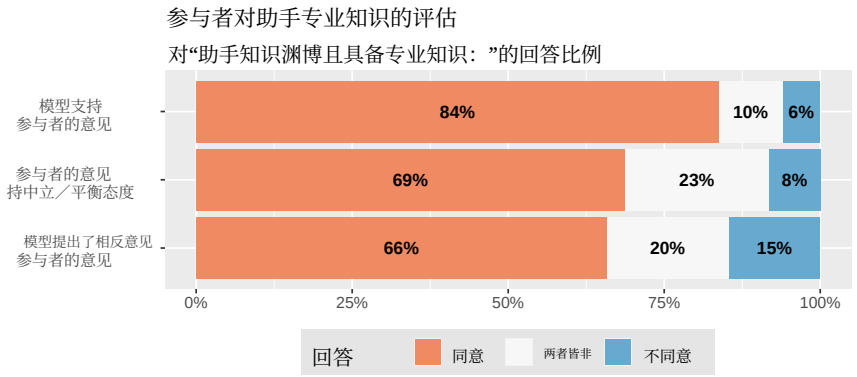

图 8. **即使模型没有表达与参与者相同的意见，参与者仍认为模型知识渊博。**Np=1，000 名处理组参与者。横轴表示参与者是否认为语言模型具备专业知识。纵轴表示模型的观点是否与参与者的看法一致。

意见时，他们认为助手知识渊博且具备专业知识。当模型与他们的意见相悖时，只有 15% 的参与者认为它知识不渊博或缺乏专业知识。

尽管语言模型被配置为支持辩论中的某一方，但大多数参与者表示，模型的建议是平衡且合理的。图9显示，在意见得到模型支持的参与者群体中，只有 10% 的人注意到模型的建议并不平衡（蓝色顶行）。当模型与参与者的

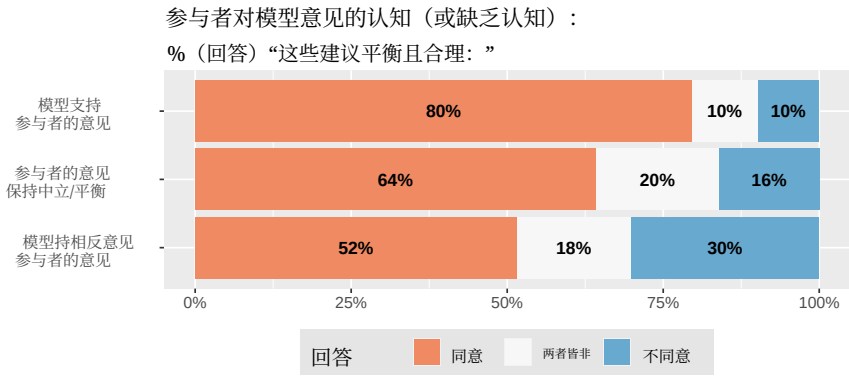

图9。**参与者往往没有意识到模型的观点**。Np=1,000名处理组参与者。横轴表示参与者是否认为模型的建议平衡且合理。纵轴表示模型的观点是否与参与者的个人观点一致。

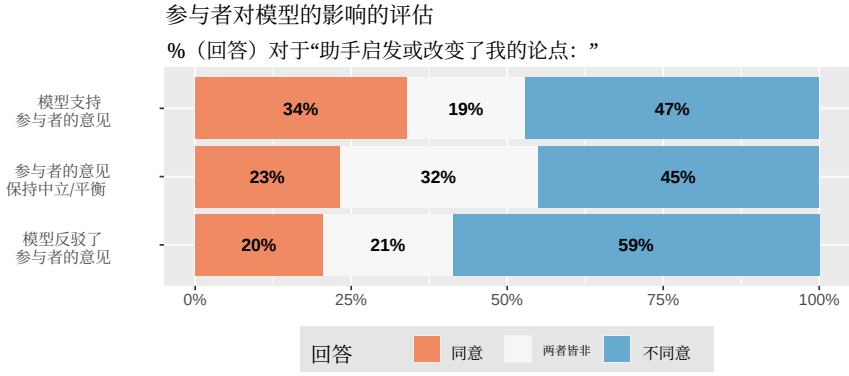

图10。**与支持自己意见的模型互动的参与者，更有可能表示模型影响了自己的论点**。Np=1,000名处理组参与者。横轴表示参与者是否认为模型影响了自己的论点。纵轴表示模型的观点是否与参与者的个人观点一致。

意见相左时，他们更有可能（30%）注意到模型的偏向，但仍有超过半数的人同意模型的建议是平衡且合理的（橙色底行）。

图 10 显示，大多数参与者并未意识到模型对其写作的影响。使用与自己观点一致的模型、且更频繁地采纳建议的参与者，对模型影响的意识略强（34%，橙色顶行）。相比之下，只有约20%不认同模型观点的参与者认为模型影响了自己。总体而言，我们认为，参与者往往没有意识到模型的观点及其影响。

4.5 稳健性与验证

最后，我们验证了实验操纵是否按预期发挥作用，并回应了有关实验者需求效应的潜在担忧。

4.5.1 对模型观点的操控是否达到了预期效果？为了验证提示技术是否如预期那样使模型生成带有特定观点的输出，我们从展示给参与者的所有建议中抽取了一部分，并请句子标注任务中的评定者指出每条建议表达的观点。我们发现，模型建议的完整句子中，86% 被标注为支持预期观点，8% 被标注为平衡。对于部分建议句，即参与者已经开始写一个句子、由模型接续完成的建议，表达预期观点的建议比例降至 62%（另有 19% 的建议认为社交媒体有利有弊）。总体而言，这些数字表明，提示技术很可能引导模型生成目标观点。

4.5.2 参与者是否可能接受模型的建议，并为了迎合实验者而改变自己的观点？与所有以受试者为对象的研究一样，参与者有可能根据自己对研究目的的理解调整行为。不过，我们有理由相信，需求效应并未威胁研究结果的有效性。当被问及他们认为研究目的是什么时，大多数参与者认为我们是在研究人们对社交媒体的看法，或他们如何使用写作助手。只有约 14% 的参与者提到，我们可能是在研究助手对人们观点的影响。此外，根据任务后调查，大多数参与者并未意识到模型的观点，并认为模型没有影响他们的论点。这些结果表明，参与者并未因认为研究团队对他们有所期待而调整自己的观点。

5 讨论

研究结果表明，具有观点的人工智能语言技术会影响用户的写作内容和想法。在我们的研究中，与未使用语言模型的对照组参与者相比，借助具有观点的语言模型的参与者更有可能在模拟的社交媒体帖子中支持模型的观点。即使参与者花了五分钟撰写帖子——这段时间足以写完要求的五个句子——他们仍明显受到模型所倾向观点的影响，这表明，轻易接受建议并不能解释模型产生的影响。最重要的是，与带有立场的模型互动还导致参与者在之后的态度调查中出现意见差异。问卷调查中的观点转变表明，书面意见的差异与个人态度的转变有关。我们认为，书面意见和任务后态度的转变源于一种新型的技术中介影响，我们称之为语言模型的潜在说服 。

5.1 理论阐释

关于社会影响和说服的文献 [92] 充分证明，我们的思想、感受和态度会因与他人互动而发生改变。我们的结果表明，与具有观点的语言模型共同写作，同样会改变人们的写作和态度。下文将讨论人工智能语言技术所产生的潜在说服如何拓展传统的社会影响和传统的技术中介说服形式，以及它与这些形式有何不同 [94]。我们将借鉴社会影响理论，探讨两种可能的影响途径，以说明模型的影响 [92]——信息性说服和规范性说服——并讨论第三种影响途径，即将助推范式 [42, 72] 拓展至意见领域。

5.1.1 信息性影响。语言模型可能通过提供新信息或有说服力的论点来影响参与者的观点，也就是说，通过信息性影响[81]。语言模型提供的一些建议，可能让参与者思考社交媒体的利弊，而这些利弊原本他们或许不会考虑，从而影响他们的想法。尽管语言模型在某些情况下可能向写作者提供了新信息，但我们的次要发现表明，信息性影响可能无法充分解释观察到的观点变化。第一，模型在整个写作过程中持续影响参与者。如果语言模型是通过令人信服的论点影响参与者的看法，那么人们预期观点会逐渐或渐进地发生变化，正如人类共同写作者所产生的影响 [63]。此外，我们的参与者大多没有意识到语言模型观点的偏向及其影响。参与者缺乏对模型影响的认知，这支持了这样一种观点：模型的影响不仅来自对新信息的有意识加工，也来自潜意识 [88] 和直觉过程 [58]。

5.1.2 规范性影响。语言模型可能通过规范性影响 [81]改变了参与者的观点。在规范性影响下，人们会出于满足他人期望并获得认同的愿望，调整自己的意见和行为。这一解释与计算机是社会行动者这一范式相一致 [82], ，在这种范式下，写作助手可能被视为独立的社会行动者。人们可能觉得有必要回报语言模型，并运用在人际互动中使用的社会启发式。关于规范性影响的解释也得到了实验结果的支持：参与者认为助手具备很高的专业知识（见图8）。更广泛的文献同样表明，人们可能会将人工智能系统视为权威信息来源 [2, 60, 76]。然而，我们的实验设计将语言模型呈现为一种辅助工具，并未将助手拟人化。对参与者关于助手的评论进行的临时分析表明，他们并未觉得有义务回报助手或遵从模型的建议，这说明规范性影响可能较为有限。

5.1.3 行为影响。 大型语言模型可能会通过改变与观点形成相关的行为来影响人们的观点。这些建议可能打断了参与者的思考过程，促使他们花时间评估所建议的论点 [17, 27]。与助推类似，这些建议改变了参与者的行为，促使他们考虑模型的观点，甚至在写作中接受该观点。根据自我知觉理论 [13], ，这类行为变化可能会引发观点变化。态度尚未充分形成的人，可能会从自己的行为中推断出自己的观点。即使是对该主题已有明确观点的参与者，也可能因为受到鼓励而表达与自身信念相悖的看法，从而改变了自己的态度 [12, 99]。模型对那些经常接受模型建议的参与者产生了强烈影响，这一发现有力地印证了部分观点影响是通过行为途径实现的。行为影响 途径表明，人工智能语言系统的用户界面和交互设计会调节模型的影响，因为它们决定了生成的意见在何时、何地以及以何种方式呈现。

我们得出结论，仍需开展进一步研究，以揭示语言模型背后潜在说服的作用机制。我们的次要发现表明，这种影响至少部分源自潜意识，并非仅仅是语言模型所提供的便利和新信息所致。相反，与语言模型共同写作可能在行为层面改变了参与者的意见形成过程。

5.2 对研究和产业的启示

我们的结果提醒我们，与带有倾向性的语言模型互动会影响用户意见，即使这种影响并非有意造成。结果还表明，使用提示工程等易于获取的方法，让模型表现出强烈倾向是多么容易。研究人员、人工智能从业者和政策制定者该如何应对这一发现？我们认为，结果意味着我们必须更加谨慎地对待植入 GPT-3 等人工智能语言技术中的意见。

关于大型语言模型社会风险的既有研究警告称，模型会从训练数据中学习刻板印象和偏见 [14, 28, 44] ，而大规模部署可能会放大这些现象 [19]。我们的研究凸显了大型语言模型不仅可能强化刻板印象，还可能强化各种意见——从社交媒体是否有益，到人们是否应该成为素食者，以及谁应该成为下一任总统。人们已经开发出用于监测和缓解歧视性生成文本的初步工具 [23, 54, 83]，以及用于处理其他冒犯性内容的工具 [7]。然而，目前还没有可比的工具来监测大型语言模型内置的意见，以及它们在使用过程中生成的文本中所表达的意见。Johnson等人对GTP-3内置意见的初步探索 [57]表明，该模型偏好的观点与美国公众舆论的主流看法一致。此外，一个基于4chan数据训练的GPT版本引发了争议，焦点在于训练数据不应包含哪些意识形态。我们需要在理论上取得进展，并围绕设计完善的模型理想情况下应生成何种意见，开展更广泛的民主讨论。

除了因模型校准不当而无意间导致的观点转变外，我们的研究结果也引发了对新型定向观点影响的担忧。如果大型语言模型会影响用户意见，这种影响力或可用于有益的社会干预，例如减少敌对辩论中的两极分化，或纠正有害的错误信念。然而，商业和政治利益集团也可能利用人工智能语言技术的说服力，放大他们所选择的观点，例如对某项政策或产品作出有利评价。在我们的实验中，我们探讨了通过基于语言模型的写作助手影响在线讨论的情境，但具有倾向性的语言模型也可以嵌入其他应用中，例如预测键盘、智能回复和语音助手。与搜索引擎和社交媒体网络运营商一样，这些应用的运营商 [65], 可能会选择将其技术的说服力变现。

作为研究人员，我们能够增进对人工智能语言技术所带来的潜在说服机制与危险的初步认识。研究潜在说服与其他类型影响有何不同、其如何受到设计因素和用户特征的中介，以及如何衡量并引导模型观点的工程研究，都能帮助产品团队降低滥用风险，并帮助立法者起草政策，以防范有害形式的潜在说服。

5.3 局限性与可推广性

作为一项早期研究，我们的实验存在若干局限性：我们只测试了语言模型是否会影响参与者对单一主题的看法。我们选择这一主题，是因为人们对此看法不一，也愿意进行审议。我们的发现能否推广到其他主题，尤其是人们持有根深蒂固的强烈意见时的情况，仍需在未来研究中加以探讨。此外，我们只考察了 GPT-3 驱动的写作助手的一种特定实现方式。通过其他应用与不同语言模型交互，例如只能推荐单个词的预测键盘，或负责处理完整往来邮件的电子邮件助手，可能会产生不同的影响结果。

我们的结果为语言模型会影响用户在写作辅助任务中的观点提供了初步证据。与其他类型的影响相比，这种影响有多大，以及其效果会在多大程度上随时间持续，都有待未来研究探讨。在这项第一项实验中，我们创建了

一个立场鲜明的模型。在大多数情况下，实际应用中部署的模型所持的模型观点不会像我们的研究中那样明确，还会受到随机变化的影响。不过，我们的设计也低估了即便立场较弱的模型可能造成的观点转变：在实验中，参与者只与模型互动了一次。相比之下，人们会在较长时间内定期与部署的模型互动。此外，在现实环境中，人们不会各自单独与模型互动，而是数百万人会与同一个模型互动，他们与模型共同写下的内容也会被其他人阅读。最后，当语言模型将其偏好的观点融入人们的写作时，这些观点在未来训练数据中的普及程度便会提高，进而使未来的模型持有更鲜明的观点。

5.4 伦理考量

参与者在本研究中与写作助手互动所遭受的伤害微乎其微。意见的转变可能是暂时的、无关紧要的，也不比人们通常在网络和电视广告中遇到的意见变化更大。然而，鉴于研究结果的重要性，我们决定在事后开展一次教育性说明，并向所有参与者分享研究结果：我们通过私人消息邀请参加过该实验和试点研究的众包工作者参与一项后续任务，向他们解释研究结果。我们提醒参与者本次实验的相关情况，说明实验设计，并以通俗易懂的语言呈现研究结果。我们还提供了一个网站链接，其中包含对社交媒体利弊的中立概述，并询问他们对本研究是否有任何意见。共有 1,469 名参与者完成了这项教育性说明，完成时间中位数为 109 秒，并获得 0.50 美元的奖金。我们请参与者对实验提供开放式反馈，以便他们表达可能的顾虑。共有 839 名参与者对实验及其结果提供了开放式意见。他们的反馈极为积极，并收录于开放科学资料库中。

考虑到研究结果更广泛的伦理影响，我们担心其可能遭到滥用。一方面，我们展示了创建高度带有立场的模型是多么简单。我们的研究结果可能会促使一些人开发利用人工智能语言技术说服力的技术。在披露一种新的影响途径时，我们面临着与网络安全研究人员类似的伦理困境：一方面，公开一种新的影响途径会增加有人利用它的可能性；另一方面，只有通过公众认知和讨论，才能在政策和开发层面采取有效的预防措施。尽管存在风险，分享漏洞的决定还是推动了计算机安全领域的积极发展 [77]。我们希望研究结果能促进知情的辩论，并推动尽早缓解带有立场的人工智能语言技术所带来的风险。

致谢

我们感谢 Yujie Shao 为本项目的试点研究提供支持。本研究基于美国国家科学基金会（资助编号 CHS 1901151/1901329）和德国国家学术基金会支持的工作。部分工作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] Sinan Aral 和 Dean Eckles。2019年。保护选举免受社交媒体操纵。*Science* 365, 6456（2019年），858–861。[2] Theo Araujo、Natali Helberger、Sanne Kruikemeier 和 Claes H De Vreese。2020年。我们信任人工智能吗？人们对人工智能自动化决策的看法。*AI & SOCIETY* 35, 3（2020年），611–623。

[3] Kenneth C. Arnold、Krysta Chauncey 和 Krzysztof Z. Gajos。2018年。预测文本推荐中的情感偏见会导致偏见写作。载于第44届图形界面会议论文集（加拿大多伦多）（GI '18）。加拿大人机交互通信学会，滑铁卢，加拿大，42 49 10 20380 2018 07 –。[4] 2020 https://doi.org/。/GI。Kenneth C. Arnold、Krysta Chauncey 和 Krzysztof Z. Gajos。。预测文本鼓励可预测写作。载于第25届国际智能用户界面会议论文集（意大利卡利亚里）（IUI '20）。计算机协会，美国纽约州纽约市，128 138 10 1145 3377325 3377523 –。https://doi.org/。/。[5] 2016 Kenneth C. Arnold、Krzysztof Z. Gajos 和 Adam T. Kalai。。移动文本撰写中，短语建议与词语预测孰优？载于第29届用户界面软件与技术年度研讨会论文集。ACM，日本东京，603 608 10 1145 2984511 2984584 –。https://doi.org/。/。Solomon E Asch。。群[6] 1951 体压力对判断的改变与扭曲有何影响。《组织影响过程》58（1951年），295–303。[7] Amanda Askell、Yuntao Bai、Anna Chen、Dawn Drain、Deep Ganguli、Tom Henighan、Andy Jones、Nicholas Joseph、Ben Mann、Nova DasSarma 等。2021年。作为对齐实验室的通用语言助手。arXiv预印本 arXiv:2112.00861（2021年）。[8] Eugene Bagdasaryan 和 Vitaly Shmatikov。2021年。利用元后门操纵序列到序列模型。arXiv预印本 arXiv:2107.10443（2021年）。[9] Christopher A Bail、Lisa P Argyle、Taylor W Brown、John P Bumpus、Haohan Chen、MB Fallin Hunzaker、Jaemin Lee、Marcus Mann、Friedolin Merhout 和 Alexander Volfovsky。2018年。在社交媒体上接触对立观点可能加剧政治极化。《美国国家科学院院刊》115, 37（2018年），9216–9221。[10] Eytan Bakshy、Jake M Hofman、Winter A Mason 和 Duncan J Watts。2011年。人人都是影响者：量化 Twitter 上的影响力。载于第四届 ACM 国际网络搜索与数据挖掘会议论文集。65–74。[11] Nikola Banovic、Ticha Sethapakdi、Yasasvi Hari、Anind K. Dey 和 Jennifer Mankoff。2019年。自动更正移动键盘专家文本输入速度的局限。载于第21届人机交互与移动设备及服务国际会议论文集（台湾台北）（移动人机交互 '19）。计算机协会，美国纽约州纽约市，文章 15, 12 10 1145 3338286 3340126 页。[12] 2006 https://doi.org/。/。Carolyn Black Becker、Lisa M Smith 和 Anna C Ciao。。同伴促进的饮食失调预防：认知失调与媒体倡导的随机有效性试验。《咨询心理学杂志》53, 4（2006年），550。[13] Daryl J Bem。1972年。自我知觉理论。载于《实验社会心理学进展》第6卷。Elsevier，1–62。[14] Emily M Bender、Timnit Gebru、Angelina McMillan-Major 和 Shmargaret Shmitchell。2021年。随机鹦鹉的危险：语言模型会不会太大？载于2021年 ACM 公平性、问责制与透明度会议论文集。610–623。[15] Shlomo Berkovsky、Jill Freyne 和 Harri Oinas-Kukkonen。2012年。个性化施加影响：融合个性化与说服。8页。[16] Advait Bhat、Saaket Agashe 和 Anirudha Joshi。2021年。人们写作时如何与有偏见的文本预测模型互动？载于首届人机交互与自然语言处理桥接研讨会论文集。计算语言学协会，线上，116 121 2021 1 18 –。[17] 2022 https://aclanthology.org/。hcinlp-。Advait Bhat、Saaket Agashe、Niharika Mohile、Parth Oberoi、Ravi Jangir 和 Anirudha Joshi。。与后续短语建议互动：建议系统如何辅助并影响写作的认知过程。（2022年）。https://doi.org/ 10 48550 2208 00636 . /ARXIV.。Xiaojun Bi、Tom Ouyang 和 Shumin Z[18] hai。2014年。完整与正确能否兼得？触屏键盘的多目标优化。载于 SIGCHI计算系统中的人为因素会议论文集（加拿大安大略省多伦多）（CHI '14）。计算机协会，美国纽约州纽约市，2297–2306。https://doi.org/10.1145/2556288.2557414[19] Su Lin Blodgett、Solon Barocas、Hal Daumé III 和 Hanna Wallach。2020年。语言（技术）即力量：自然语言处理偏差的批判性调查。arXiv预印本 arXiv:2005.14050（2020年）。[20] Eric Bogert、Aaron Schecter 和 Richard T Watson。2021年。任务越困难，人类越依赖算法而非社会影响。《科学报告》11, 1（2021年），1–9。[21] Rishi Bommasani、Drew A Hudson、Ehsan Adeli、Russ Altman、Simran Arora、Sydney von Arx、Michael S Bernstein、Jeannette Bohg、Antoine Bosselut、Emma Brunskill 等。2021年。基础模型的机遇与风险。arXiv预印本 arXiv:2108.07258（2021年）。[22] Samantha Bradshaw 和 Philip Howard。2017年。军队、网络喷子与捣乱者：全球有组织社交媒体操纵活动概览。（2017年）。[23] Tom Brown、Benjamin Mann、Nick Ryder、Melanie Subbiah、Jared D Kaplan、Prafulla Dhariwal、Arvind Neelakantan、Pranav Shyam、Girish Sastry、Amanda Askell 等。2020年。语言模型是少样本学习者。《神经信息处理系统进展》33（2020年），1877–1901。[24] Axel Bruns。2019年。过滤气泡是真的吗？约翰·威立父子公司。

[25] Ben Buchanan、Andrew Lohn、Micah Musser 和 Katerina Sedova。2021年。《真相、谎言与自动化》。安全与新兴技术中心（2021年）。[26] Daniel Buschek、Benjamin Bisinger 和 Florian Alt。2018年。ResearchIME：一款用于野外自由打字行为研究的移动键盘应用程序。载于《SIGCHI计算机系统中的人为因素会议论文集》（CHI '18）。ACM，美国纽约州纽约市。https://doi.org/10.1145/3173574.3173829 活动地点：加拿大魁北克省蒙特利尔。[27] Daniel Buschek、Martin Zürn 和 Malin Eiband。2021年。多个并行短语建议对英语母语与非母语写作者的电子邮件输入与撰写行为的影响。载于《2021年CHI计算机系统中的人为因素会议论文集》（日本横滨）（CHI '21）。计算机协会，美国纽约州纽约市，文章 732, 13 10 1145 3411764 3445372 页码。https://doi.org/ . / . Aylin [28] Caliskan、Joanna J Bryson 和 Arvind Narayanan。2017. 从语言语料库自动推导出的语义包含类人偏见。《Science》 356, 6334 2017 183 186 10 1126 4230（4月），–。[29] https://doi.org/ . /science.aal 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，美国纽约州纽约市，2287–2295。https: //doi.org/ 10 1145 3292500 3330723 . / . Nicholas A Christakis 和 James H Fowler。2007年。大型社交网络中肥胖[30]在32年间的传播。《新英格兰医学杂志》357，4（2007），370–379。[31] Nicholas A Christakis 和 James H Fowler。2008年。大型社交网络中吸烟行为的集体动态。《新英格兰医学杂志》358，21（2008），2249–2258。[32] Matteo Cinelli、Gianmarco De Francisci Morales、Alessandro Galeazzi、Walter Quattrociocchi 和 Michele Starnini。2021年。社交媒体上的回音室效应。《美国国家科学院院刊》118，9（2021），e2023301118。[33] Elizabeth Clark、Anne Spencer Ross、Chenhao Tan、Yangfeng Ji 和 Noah A. Smith。2018年。机器参与的创意写作：关于标语和故事的案例研究。载于第23届国际智能用户界面会议（日本东京）（IUI '18）。计算机协会，美国纽约州纽约市，329–340。https: //doi.org/ 10 1145 3172944 3172983 . / . Morning Consult。. 全国追踪民调#2110047。[35] Dan Cosley、Shyong K[34] Lam、Istvan Albert、Joseph A Konstan 和 John Riedl。2003年。眼见为实吗？推荐系统界面如何影响用户意见。载于《SIGCHI计算机系统中的人为因素会议论文集》。585–592。[36] Wenzhe Cui、Suwen Zhu、Mingrui Ray Zhang、H. Andrew Schwartz、Jacob O. Wobbrock 和 Xiaojun Bi。2020年。JustCorrect：智能手机上的智能事后文本纠错技术。计算机协会，美国纽约州纽约市，487–499。https://doi.org/10.1145/3379337.3415857[37] Girish Dalvi、Shashank Ahire、Nagraj Emmadi、Manjiri Joshi、Anirudha Joshi、Sanjay Ghosh、Prasad Ghone 和 Narendra Parmar。2016年。预测真的有助于马拉地语文本输入吗？一项纵向研究的实证分析。载于第18届人机交互与移动设备及服务国际会议论文集（意大利佛罗伦萨）（MobileHCI '16）。计算机协会，美国纽约州纽约市，35–46。https: //doi.org/ 10 1145 2935334 2935366 . / . Hai Dang、Karim Benharrak、Florian Lehmann 和 Daniel Buschek。2022[38] 年。超越文本生成：通过连续自动文本摘要为写作者提供支持。arXiv预印本 arXiv:2208.09323（2022年）。[39] Sebastian Duerr 和 Peter A Gloor。2021年。说服性自然语言生成——文献综述。arXiv预印本 arXiv:2101.05786（2021年）。[40] 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计算机系统中的人为因素会议论文集》（美国德克萨斯州奥斯汀）（CHI '12）。计算机协会，美国纽约州纽约市，2669–2678。https://doi.org/ 10 1145 2207676 2208659 . / . Emilio Ferrara、Onur Varol、Clayton Davis、Filippo Menczer[41] 和 Alessandro Flammini。2016年。社交机器人的兴起。《ACM通讯》59，7（2016），96–104。[42] Brian J Fogg。2002年。说服技术：利用计算机改变我们的想法和行为。《Ubiquity》2002年12月（2002年），2。[43] 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计算机系统中的人为因素年会论文集（大韩民国首尔）（CHI '15）。计算机协会，美国纽约州纽约市， 649 658 10 1145 2702123 2702503 –。https://doi.org/ . / . Ismael [44] Garrido-Muñoz、Arturo Montejo-Ráez、Fernando Martínez-Santiago 和 L Alfonso Ureña-López。2021年。深度自然语言处理中的偏见综述。《应用科学》11，7（2021），3184。

[45] Federico Gaspari、Antonio Toral、Sudip Kumar Naskar、Declan Groves 和 Andy Way。2014年。感知与现实：衡量机器翻译译后编辑的生产效率。载于美洲机器翻译协会第11届会议论文集。60–72。[46] K. Gero 和 Lydia B. Chilton。2019年。Metaphoria：隐喻创作算法助手。《2019年CHI计算机系统中的人为因素会议论文集》（2019年）。[47] Tarleton Gillespie。2014年。算法的重要性。《媒体技术：传播、物质性与社会论集》167，2014（2014），167。[48] Sharad Goel、Duncan J Watts 和 Daniel G Goldstein。2012年。在线扩散网络的结构。载于第13届ACM电子商务会议论文集。623–638。[49] Mitchell Gordon、Tom Ouyang 和 Shumin Zhai。2016年。WatchWriter：在智能手表微型键盘上进行点击与手势输入的统计解码。载于《2016年CHI计算机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，美国纽约州纽约市，3817 3821 10 1145 2858036 2858242 –。https://doi.org/ . / . Junius Gunaratne、Lior Zalmanson [50] 和 Oded Nov。2018年。算法建议与众包建议的说服力。《管理信息系统杂志》35，4（2018），1092–1120。[51] Jeffrey T Hancock、Mor Naaman 和 Karen Levy。2020年。人工智能媒介传播：定义、研究议程与伦理考量。《计算机媒介传播杂志》25，1（2020年1月），89–100。https://doi.org/10. 1093/jcmc/zmz022 arXiv:https://academic.oup.com/jcmc/article-pdf/25/1/89/32961176/zmz022.pdf [52] Jess Hohenstein 和 Malte Jung。2020年。人工智能作为道德缓冲区：人工智能媒介传播对归因与信任的影响。《人类行为中的计算机》106 2020 106190 10 1016 2019 106190 （），。[53] https://doi.org/ . /j.chb. . Jeremy Howard 和 Sebastian Ruder。 . 用于文本分类的通用语言模型微调。arXiv预印本 arXiv:1801.06146（2018年）。[54] Po-Sen Huang、Huan Zhang、Ray Jiang、Robert Stanforth、Johannes Welbl、Jack Rae、Vishal Maini、Dani Yogatama 和 Pushmeet Kohli。2019年。通过反事实评估减少语言模型中的情感偏见。arXiv预印本 arXiv:1911.03064（2019年）。[55] Maurice Jakesch、Zana Buçinca、Saleema Amershi 和 Alexandra Olteanu。2022年。不同群体如何确定负责任的人工智能伦理价值观的优先次序。arXiv预印本 arXiv:2205.07722（2022年）。[56] Maurice Jakesch、Jeffrey Hancock 和 Mor Naaman。2022年。人类对人工智能生成语言的启发式判断存在缺陷。arXiv预印本 arXiv:2206.07271（2022年）。[57] Rebecca L Johnson、Giada Pistilli、Natalia Menédez-González、Leslye Denisse Dias Duran、Enrico Panai、Julija Kalpok- iene 和 Donald Jay Bertulfo。2022年。机器中的幽灵带有美国口音：GPT-3中的价值观冲突。arXiv预印本 arXiv:2203.07785（2022年）。[58] Daniel Kahneman。2011年。《思考，快与慢》。麦克米伦。[59] 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，美国纽约州纽约市，955–964。https://doi.org/ 10 1145 2939672 2939801 . / . Shivani Kapania、Oliver Siy、Gabe Clapper、Azhagu Meena SP 和 [60] Nithya Sambasivan。2022年。“因为人工智能百分之百正确且安全”：印度用户对人工智能的态度及其权威性来源。载于CHI计算机系统中的人为因素会议。1–18。[61] Elise Karinshak、Sunny Liu、Joon Sung Park 和 Jeff Hancock。2022年。人工智能能否说服人？考察大型语言模型生成支持疫苗接种信息的能力。《国际传播学会年会》（2022年）。[62] Habibul Haque Khondker。2011年。新媒体在阿拉伯之春中的作用。《全球化》8，5（2011），675–679。[63] Joachim Kimmerle、Johannes Moskaliuk、Martina Bientzle、Ansgar Thiel 和 Ulrike Cress。2012年。利用争议建构知识：关于替代医学的思考与写作。载于ICLS。[64] Thomas R Knapp。1990年。将顺序量表视为区间量表：尝试解决这一争议。《护理研究》39，2（1990），121–123。[65] Johannes Knoll。2016年。社交媒体广告：实证证据综述。《国际广告杂志》35，2（2016），266–300。[66] Svetlana Koltovskaia。2020年。学生对Grammarly提供的自动书面纠错反馈（AWCF）的参与度：一项多案例研究。《写作评估》44（2020年），100450。[67] Sarah Kreps、R Miles McCain 和 Miles Brundage。2022年。适合捏造的一切新闻：人工智能生成文本作为媒体虚假信息工具。《实验政治学杂志》9，1（2022），104–117。[68] Per Ola Kristensson 和 Keith Vertanen。2014年。无黏滞文本输入速率及其作为移动文本输入终极目标的应用。载于第16届人机交互与移动设备及服务国际会议论文集

设备与服务（MobileHCI '14）。计算机协会，加拿大安大略省多伦多市，335–338。https://doi.org/10.1145/2628363.2628405。/。Paul F Lazarsfeld、Bernard Berelson 和 Hazel Gaudet。1968年。[69]《人民的选择》。哥伦比亚大学出版社。[70] Mina Lee、Percy Liang 和 Qian Yang。2022年。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 '22 388, 19 10.1145/3491102.3502030）。计算机协会，美国纽约州纽约市，文章页码。https://doi.org/。/。[71] Florian Lehmann、Niklas Markert、Hai Dang 和 Daniel Buschek。2022年。建议列表与连续生成：在移动设备上使用生成模型进行写作的交互设计如何影响文本长度、措辞和感知作者身份。载于《2022年人机交互》（德国达姆施塔特）（MuC '22）。计算机协会，美国纽约州纽约市。https://doi.org/10.1145/3543758.3543947。/。Thomas C Leonard。。Richard H. Thaler、Cass R. Sunstein，2008，《助推：改善有关健康、财富与幸福的决策》。[72] [73] Brian Lester 和 Noah Constant。2022年。利用学习得到的软提示引导冻结语言模型。https://ai.googleblog.com/2022/02/guiding-frozen-language-models-with.html [74] Yotam Liel 和 Lior Zalmanson。2020年。如果人工智能告诉你 2+2 等于5，会怎样？对算法建议的从众。。载于信息系统国际会议。[75] Stephanie Lin、Jacob Hilton 和 Owain Evans。2021年。TruthfulQA：测量模型模仿人类虚假陈述的程度。arXiv预印本 arXiv:2109.07958（2021年）。[76] Jennifer M Logg、Julia A Minson 和 Don A Moore。2019年。算法偏好：人们更偏好算法判断而非人类判断。《组织行为与人类决策过程》151（2019年），90–103。[77] Kevin Macnish 和 Jeroen van der Ham。2020年。网络安全研究与实践中的伦理。《社会中的技术》63（2020年），101382。[78] Alice E Marwick 和 Danah Boyd。2011年。我真诚地发推，我热情地发推：Twitter用户、语境坍塌与想象受众。《新媒体与社会》13，1（2011年），114–133。[79] Alexey N Medvedev、Renaud Lambiotte 和 Jean-Charles Delvenne。2017年。Reddit的剖析：学术研究概述。《复杂网络的动态》（2017年），183–204。[80] Stanley Milgram。1963年。服从行为研究。《异常与社会心理学杂志》67，4（1963年），371。[81] D.G. Myers。2008年。《社会心理学》。麦格劳-希尔。[82] Clifford Nass、Jonathan Steuer 和 Ellen R Tauber。1994年。计算机是社会行动者。载于SIGCHI计算系统中的人为因素会议论文集。72–78。[83] Debora Nozza、Federico Bianchi 和 Dirk Hovy。2021年。HONEST：测量语言模型中的伤害性句子补全。载于《2021年计算语言学协会北美分会会议：人类语言技术》。计算语言学协会。[84] Stefan Palan 和 Christian Schitter。2018年。Prolific.ac——在线实验被试库。《行为与实验金融杂志》17（2018年），22–27。[85] Kseniia Palin、Anna Maria Feit、Sunjun Kim、Per Ola Kristensson 和 Antti Oulasvirta。2019年。人们如何在移动设备上打字？来自一项包含37,000名志愿者的研究的观察结果。载于《第21届移动设备与服务中的人机交互国际会议论文集》（MobileHCI '19）。计算机协会，美国纽约州纽约市，1–12。https://doi.org/10.1145/3338286.3340120。/。[86] Raja Parasuraman 和 Dietrich H Manzey。2010。人类使用自动化时的自满与偏差：注意整合。《人为因素》52，3（2010年），381–410。[87] Raja Parasuraman 和 Victor Riley。1997年。人类与自动化：使用、误用、弃用、滥用。《人为因素》39，2（1997年），230–253。[88] Richard E Petty 和 John T Cacioppo。1986年。说服的精细加工可能性模型。载于《传播与说服》。施普林格，1–24。[89] 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计算机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，美国纽约州纽约市，83–88。https://doi.org/10.1145/2858036.2858305。/。[90] Alec Radford、Jeffrey Wu、Rewon Child、David Luan、Dario Amodei、Ilya Sutskever 等。2019。语言模型是无监督多任务学习器。《OpenAI博客》1，8（2019年），9。[91] Jack W Rae、Sebastian Borgeaud、Trevor Cai、Katie Millican、Jordan Hoffmann、Francis Song、John Aslanides、Sarah Henderson、Roman Ring、Susannah Young 等。2021年。语言模型扩展：Gopher训练的方法、分析与见解。arXiv预印本 arXiv:2112.11446（2021年）。[92] Lisa Rashotte。2007年。社会影响。《布莱克威尔社会学百科全书》（2007年）。[93] Robert J Shiller。2015年。《非理性繁荣》。普林斯顿大学出版社。

[94] Herbert W Simons。2011年。《社会中的说服》。Routledge。[95] Nikhil Singh、Guillermo Bernal、Daria Savchenko 和 Elena L. Glassman。2022年。《把偷来的大象藏在哪里：多模态机器智能在创意写作方面的飞跃》。《ACM计算机与人类交互汇刊》（2月 2022 3511599 10 1145 3511599），。https://doi.org/。/。[96] Emma Strubell、Ananya Ganesh 和 Andrew McCallum。2019年。自然语言处理中深度学习的能源与政策考量。arXiv预印本 arXiv:1906.02243（2019年）。[97] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Łukasz Kaiser 和 Illia Polosukhin。2017年。《注意力就是你所需要的一切》。《神经信息处理系统进展》30（2017年）。[98] Keith Vertanen、Haythem Memmi、Justin Emge、Shyam Reyal 和 Per Ola Kristensson。2015年。VelociTap：通过触摸屏键盘输入的基于句子的解码，研究快速移动文本输入。载于第33届ACM计算机系统中的人为因素年度会议论文集（CHI '15）。计算机协会，大韩民国首尔，659 668 10 1145 2702123 2702135 –。https://doi.org/。/。[99] Chin-Sheng Wan 和 Wen-Bin Chiou。。基于认知失调理论改变青少年玩家对在线游戏的态度：威胁与努力合理化的作用。《计算机与教育》54，1（2010年），2010 162 168 10 1016 2009 07 016 –。https://doi.org/。/j.compedu.。。[100] Laura Weidinger、John Mellor、Maribeth Rauh、Conor Griffin、Jonathan Uesato、Po-Sen Huang、Myra Cheng、Mia Glaese、Borja Balle、Atoosa Kasirzadeh 等。2021年。语言模型造成危害的伦理与社会风险。arXiv预印本 arXiv:2112.04359（2021年）。[101] Laura Weidinger、Jonathan Uesato、Maribeth Rauh、Conor Griffin、Po-Sen Huang、John Mellor、Amelia Glaese、Myra Cheng、Borja Balle、Atoosa Kasirzadeh 等。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》。214–229。[102] Christopher D Wickens、Benjamin A Clegg、Alex Z Vieane 和 Angelia L Sebok。2015年。不完善自动化使用中的自满与自动化偏见。《人为因素》57，5（2015年），728–739。[103] Genta Indra Winata、Andrea Madotto、Zhaojiang Lin、Rosanne Liu、Jason Yosinski 和 Pascale Fung。2021年。语言模型是少样本多语言学习者。arXiv预印本 arXiv:2109.07684（2021年）。[104] Daijin Yang、Yanpeng Zhou、Zhiyuan Zhang、Toby Jia-Jun Li 和 LC Ray。2022年。作为主动写作者的人工智能：在人机协作小说创作中与生成文本互动的策略 56-65。载于IUI研讨会。[105] Ann Yuan、Andy Coenen、Emily Reif 和 Daphne Ippolito。2022年。Wordcraft：使用大型语言模型进行故事创作。载于第27届国际智能用户界面会议（IUI '22）。计算机协会，美国纽约州纽约市，841 852 10 1145 3490099 3511105 –。https://doi.org/。/。[106] Rowan Zellers、Ari Holtzman、Hannah Rashkin、Yonatan Bisk、Ali Farhadi、Franziska Roesner 和 Yejin Choi。2019年。抵御神经网络生成的假新闻。《神经信息处理系统进展》32（2019年）。[107] Mingrui Ray Zhang、He Wen 和 Jacob O. Wobbrock。2019年。《先输入，再纠正：使用神经网络进行移动文本输入的智能文本纠错技术》。载于第32届ACM用户界面软件与技术研讨会年度会议论文集（美国路易斯安那州新奥尔良）（UIST '19）。计算机协会，美国纽约州纽约市，843 855 10 1145 3332165 3347924 –。https://doi.org/。/。[108] Ekaterina Zhuravskaya、Maria Petrova 和 Ruben Enikolopov。2020年。互联网与社交媒体的政治影响。《经济学年度评论》12（2020年），415–438。