在 X 上使用 Grok 纠正错误信息，同时避免人身攻击

Kevin Matthe Caramancion国防分析系 海军研究生院， 美国国防部（由美国海军管理）加利福尼亚州蒙特雷，美国 orcid.org/0000-0003-1332-3426

摘要——**在公共在线空间纠正错误信息，往往会使用户遭遇敌意和人身攻击，进而打击他们参与纠正性话语的积极性。本研究提供了实证证据，表明在 X 上调用原生大语言模型 Grok，而非直接与其他用户对峙，与错误信息纠正过程中出现的不同社会反应相关。本研究采用观察性研究设计，分析了五个冲突激烈的错误信息主题下的 100 条纠正回复，其中 Grok 介导的纠正与人工发布的直接回应各占一半。主要结果指标是纠正发布后 24 小时内是否遭到至少一次人身攻击。人工发布的纠正中有 72% 遭遇人身攻击，而 Grok 介导的纠正均未遭遇此类攻击。卡方检验证实，两者之间存在统计学显著关联，且效应量较大。这些发现表明，人工智能媒介化纠正可能通过减少回应错误信息时的人际敌意，改变公共分歧中的社会动态。**

关键词——错误信息、在线攻击、人工智能、***Grok***、社交媒体冲突、事实核查

一. 引言

在线虚假信息纠正面临的一项长期障碍，是用户在质疑虚假主张时所感知到的个人风险。在 X 等平台上，指出帖子内容不准确往往会引发人身攻击、敌意和报复行为。这种局面会阻碍用户参与公开事实核查，并削弱集体抵制欺骗性叙事的能力。先前研究表明，在线行为反应受到感知到的社会威胁影响；当用户预料到个人反噬时，往往会避免纠正错误信息 [1]。随着在线交流日益充满争议，了解个人如何应对这些风险，对于更广泛地研究错误信息管控至关重要。

X 上一种日益普遍的行为模式揭示了用户管理这些风险的新策略。如今，许多用户会调用 Grok——该平台内置的大语言模型——来针对误导性帖子生成纠正性回应。用户不再亲自发布纠正内容，而是提示 Grok 评估某项主张，再展示人工智能生成的解释。这种做法使用户能够与纠正行为保持距离，并将评估虚假内容的责任转移给 Grok。

早期研究发现，用户会根据自身的安全感以及感知到的遭受人身攻击的脆弱性，调整应对错误信息的方式，这凸显了网络纠正行为背后的社会复杂性 [1]。在此背景下，借助 Grok 可能发挥保护性缓冲的作用，使用户能够质疑错误信息，同时减少直接遭受敌意的风险。尽管关于错误信息检测、用户脆弱性以及影响网络回应的心理因素已有大量研究 [1][2]，但鲜有学术研究探讨在高度冲突的情境中，人们如何将人工智能介导的工具用作人际屏障。

现有研究主要聚焦于检测准确率、认知易感性、接触效应或一般行为模式，但尚未探讨用户如何利用人工智能模型充当社会中介，以降低个人风险。这一研究空白限制了我们对人工智能介导的传播如何重塑现实环境中的虚假信息纠正动态的理解。

本研究通过考察 X 上的用户在回应错误信息时如何调用 Grok，以及这类纠正如何在社会层面被接受，来填补这一空白。分析围绕以下两个研究问题展开：

RQ1. 在 *X* 上，使用 *Grok* 进行错误信息纠正的情况如何呈现在不同的错误信息主题中？

RQ2. 在 *X* 上回应错误信息的过程中，与人工直接纠正相比，*Grok*介导的纠正获得个人攻击或人身攻击的可能性是否存在差异？

本文的理论贡献在于整合了错误信息研究、在线行为研究与人工智能中介互动。本文通过识别人工智能系统在公共话语中的一种尚未得到充分研究的社会功能，拓展了现有文献，具体而言，即人工智能系统在容易引发冲突的纠正情境中所发挥的中介作用。实践贡献则在于揭示用户如何在充满敌意的在线空间中调整自身行为，以及平台整合的人工智能功能可能如何塑造纠正与互动的模式。

最后，本文的结构安排如下。下一节将回顾有关错误信息纠正、在线冲突行为和人工智能支持的传播的相关文献。随后一节将详细介绍本研究所采用的数据收集程序、编码步骤和分析方法。接着呈现研究发现及其解读，最后讨论研究启示与未来研究方向。

II. 文献背景

A. 在线纠正、敌意与人身攻击的社会风险

研究一再表明，在社交场合纠正错误信息会使用户面临社交和人际风险。当个人公开质疑不准确的内容时，往往会遭遇敌意或人身攻击式回应，从而打击他们今后参与纠正行为的意愿。Del Saz-Rubio（2023年）发现，在有争议的在线交流中，不礼貌和攻击性语言屡见不鲜，尤其是在政治话语中[3]。此外，Smith-Jones（2023年）记录了人们在在线对话中如何运用叙事和语言策略，在讨论敌对或敏感话题时避免直接的人际冲突，说明媒介化表达可以在争议性交流中发挥缓冲作用

中 [4]。这些动态会影响用户的预期，使人们在纠正错误信息时更加谨慎或有所选择。Ahn 等人（2015年）证明，媒介化体验能够通过降低感知到的社交和评价威胁，改变行为参与。他们的研究结果表明，相较于通过直接的人际对抗来采取行动，将纠正行为或亲社会行为置于媒介化系统中呈现时，人们会报告更高的回应效能和更强的行动意愿。这支持了一个更广泛的论点：社会评价和对负面反馈的预期会抑制公众参与，而媒介化或间接参与则能降低这些障碍，促进纠正行为[5]。

B. 大语言模型与错误信息：风险、机遇及社会技术后果

大语言模型在错误信息生态系统中扮演着复杂的角色。它们既可能扩大错误信息的传播，也能助力纠正错误信息。Chen 和 Shu（2024年）回顾了当前研究，并得出结论：大语言模型是检测错误信息的有力工具，同时也指出它们有能力生成具有说服力的虚假内容 [6]。Pan 等人（2023年）探讨了大语言模型影响的双重性质，并表明这些模型在遭到滥用时，可能会无意中助长错误信息宣传活动 [7]。

与此同时，旨在进行循证推理的模型在支持纠正错误信息方面展现出潜力。Zhou 及其同事（2024年）证明，与未经辅助的人类判断相比，在大语言模型协助下进行主张核验可以提高准确性和一致性 [8]。

C. 大语言模型作为中介与新兴的人工智能媒介化纠正

一系列新兴研究将大语言模型视为信息工具之外的社会中介。Lim 等人（2025年）将大语言模型描述为沟通中介，用户在公共平台上应对冲突和错误信息时可能会依赖它们。这一观点表明，大语言模型可以充当人际缓冲者，而不只是分类系统 [9]。

相关研究支持这一发展趋势。Gabriel 等人（2024）发现，基于大语言模型的干预可以影响在线对话的语气和走向，这表明其具备超越信息处理的人际功能 [10]。此外，Fu、Foell、Xu 和 Hiniker（2024）指出，人们会在高风险的人际情境中使用人工智能媒介化传播工具来撰写或修改消息，通常是为了调整语气、增强信心并避免潜在冲突 [11], , 这与在纠正错误信息时将 Grok 作为社交缓冲的观点相符。

D. 信任、准确性与人工智能媒介化传播面临的挑战

尽管大语言模型输出具有潜力，但人们仍对其可信度和可靠性存有顾虑。Mittelstadt 等人（2023）警告称，大语言模型经常会以令人信服的自信语气给出错误回答；若在事实争议中依赖此类系统，风险便会随之增加 [12]。艾伦·图灵研究所（2023）也提醒，如果未经谨慎评估便加以使用，大语言模型可能会在选举相关错误信息的情境中无意间加剧危害[13]。

这些担忧凸显出，不仅要考察大语言模型的准确性，也要审视它们在用户调用时所发挥的社会功能。正如 Gabriel 等人（2024年）所指出的，人工智能介导的传播可能改变人际互动，而传统的错误信息研究尚未充分探讨这些影响。

E. 文献缺口：缺乏对大语言模型作为社交屏障的研究

现有文献表明，纠正错误信息会带来社交风险，大语言模型在错误信息生态系统中发挥着日益重要的作用，而人工智能中介互动可能会重塑线上交流。然而，目前尚无研究通过实证评估，在纠正错误信息时调用人工智能模型是否与人身攻击减少有关。尽管已有关于大语言模型作为交流中介的概念性讨论，但实证证据仍然有限。本研究通过考察在 X 上高度冲突的错误信息语境中，与人工直接纠正相比，Grok介导的纠正是否伴随着不同的社会反应，来填补这一研究空白。

III. 研究方法

A. 研究设计

本研究采用观察性研究设计，通过结构化抽样程序考察 X 上自然发生的互动。本研究并未将虚假信息帖子视为主要分析单位，而是将单条纠正回复作为分析单位。这种做法使抽样策略与研究问题直接对应：研究关注的是用户在发布纠正后经历了怎样的社会反应，而非原始虚假信息帖文本身的特征。

B. 错误信息主张的选择

根据独立事实核查组织此前进行的核查，选取了五项错误信息主张。所选主张能够代表在 X 上广泛传播、反复出现且有充分记录的虚假叙事。这些主张仅用于界定数据收集的主题范围，并未被视为分析单位。

C. 错误信息帖子的识别

针对每项错误信息主张，使用关键词搜索和平台搜索工具，识别出 X 上多条公开发布且声称该主张属实的帖子。符合条件的帖子须可公开访问、包含目标主张，且至少有一条回复。此阶段未设置互动量门槛，以避免优先选择病毒式传播或被大幅放大的内容。这些帖子仅作为提取纠正回复的来源。

D. 纠正回复的提取

所有符合条件的虚假信息帖子下的回复均经过人工审核，以识别纠正回复。如果某条回复明确质疑原始说法的准确性、提供事实性反驳信息，或将该说法标记为虚假或具有误导性，就会被归类为纠正。随后，纠正被分为两类。直接纠正是指完全以用户自己的口吻撰写，且未提及人工智能系统的纠正。Grok中介纠正是指明确提及 Grok、引用 Grok 的评估，或在平台界面中展示明显由 Grok 生成文本的纠正。该类别还包括明确提示 Grok 评估某项说法真实性的回复，例如询问 Grok 某项说法是否属实，前提是该提示明确处于虚假信息纠正的语境中

，并引出或预期得到评价性回应 [10]。

E. 抽样策略与选择偏差的降低

为减少与帖子可见度、回复排序和平台排名算法相关的选择偏差，我们在纠正回复层面而非虚假信息帖子层面进行随机抽样。针对每项错误信息主张，将所有符合条件的纠正回复按纠正类型汇总。从每个汇总组中，使用随机数生成器随机抽取十条直接纠正和十条Grok中介纠正。最终共获得100条纠正回复的样本，其中包括50条直接纠正和50条Grok中介纠正。通过对纠正而非帖子进行抽样，无论原始讨论串的互动程度或显著性如何，每条符合条件的纠正都有相同的入选概率。

F. 结果测量与编码程序

主要因变量是纠正回复是否至少收到一条人身攻击。对于每条抽样得到的纠正，我们收集了固定的24小时观察窗口内针对该纠正的所有回复。人身攻击被定义为针对纠正者本人，而非针对纠正内容作出回应。每条纠正均采用二分法编码：出现至少一条人身攻击回复时记为1，否则记为0。

G. 分析方法

为回答第一个研究问题，本研究采用描述性分析，记录多个错误信息主题中是否出现 Grok 介导的纠正，而非估算其在更广泛平台上的普遍程度。由于纠正类型在研究设计中经过均衡，因此未对总体频率作出推断。

为回答第二个研究问题，本研究采用卡方独立性检验，考察纠正类型（Grok 介导的纠正与人工直接纠正）和人身攻击发生情况之间的关联。根据在既定观察窗口内是否观察到至少一条人身攻击回复，对因变量进行二分编码。效应量采用 φ 系数评估。所有分析均以纠正为单位进行。

H. 人身攻击的定义

在本研究中，若回复针对某项纠正，并针对纠正者本人而非纠正内容进行攻击，则视为发生了人身攻击。这包括侮辱、辱骂、贬损言论、针对智力或品格的指责，以及轻蔑的针对个人的标签。批评论点、质疑证据，或仅表示不同意但未针对个人的回复，不编码为人身攻击。若在既定观察窗口内出现至少一条此类回复，则将该项纠正编码为遭到人身攻击。

I. 伦理考量

所有数据均收集自 X 上公开访问的帖子。分析和报告过程中，用户标识符均已匿名化。研究未尝试与用户互动，也未试图影响平台行为。

J. 局限性

本研究存在若干局限性。首先，分析依据的是公开可观察到的行为，无法获知用户的内在意图或动机。尽管研究根据对 Grok 的明确提及或可识别的 AI 生成内容来判定 Grok中介纠正，但无法确定用户是否有意借助 Grok 来避免人身攻击，或是出于便利等其他原因。因此，研究结果反映的是可观察到的纠正行为与社会反应之间的关联，而非用户意图的直接证据。其次，本研究考察的错误信息主张和纠正回复数量有限，这可能会限制研究结果对其他主题或平台的普适性。最后，由于数据属于观察性数据且仅来自单一平台，因此无法确凿地得出关于调用 Grok 对社会反应影响的因果结论。

K. 搜索程序

使用 X 的搜索界面和网络搜索工具（例如网站限定查询 site:x.com）进行搜索。在可能的情况下，使用已退出登录的浏览器会话进行搜索，以减少个性化因素的影响。所有搜索均统一使用相同的搜索词和时间范围。

L. 纳入标准

符合以下所有条件的更正回复将被纳入：

1. 它回应了一项经独立事实核查组织核实为虚假的说法 2. 它对该说法作出了纠正性回应 3. 它可公开访问 4. 至少有一条回复是针对它的

M. 排除标准

符合以下情况的纠正性回复将被排除：

1. 内容明显带有讽刺或戏仿性质 2. 内容来自自动化账号或垃圾账号 3. 内容提及 Grok，但未对说法进行评估 4. 内容缺乏明确可辨的错误信息背景

IV. 研究发现与结果

A. 样本概览

最终数据集由从 X 收集的 100 条虚假信息纠正回复组成，涉及五个错误信息主题：否认选举结果、未经证实的疫苗主张、移民与犯罪叙事、否认气候变化，以及声称重大大规模枪击事件是策划或捏造的。每个主题分析了 20 条纠正回复，其中 Grok 介导的纠正与人工直接纠正各占一半。最终样本包括 50 条 Grok介导的纠正和 50 条人工发布的纠正。

B. 描述性结果

在全部五个错误信息主题中，50条人工发布的纠正里有36条出现了人身攻击，

相当于72%的人工纠正。相比之下，50条Grok介导的纠正均未遭到人身攻击。在所考察的所有错误信息主题中，这一模式始终一致。

按主题来看，针对否认选举结果说法的人工发布的纠正中，有 8 例（共 10 例）遭到人身攻击式回应；针对否认气候变化说法的纠正中，有 6 例；针对未经证实的疫苗主张的纠正中，有 9 例；针对移民与犯罪叙事的纠正中，有 6 例；针对大规模枪击事件骗局说法的纠正中，有 7 例。五个主题中，Grok介导的纠正均未遭到人身攻击式回应。

C. 推断分析

我们进行了卡方独立性检验，以考察纠正类型与是否发生人身攻击之间的关联。分析发现，纠正类型与人身攻击的发生之间存在统计学上的显著关联，$\chi 2(1, N = 100) = 56.25$，$p < .001$。φ系数为 .75，表明效应量较大。这些结果表明，与人工直接纠正相比，Grok介导的纠正遭到人身攻击的可能性显著较低。

D. 表格与图示引用

表1展示了所有错误信息主题中按纠正类型划分的人身攻击分布。

表 I.　　按纠正类型划分的人身攻击分布（汇总所有主题）

纠正类型	人身攻击 Yes	人身攻击 No	总计
人工	36	14	50
经 Grok 处理	0	50	50
总计	36	64	100

注：每一行代表一条独立的错误信息纠正回复。若回复中至少出现一处针对纠正者个人或品格的攻击，则将其编码为人身攻击。

表 2 按纠正类型对人身攻击结果进行了主题层面的细分。

表 II.　　按错误信息主题和纠正类型划分的人身攻击结果

错误信息主题	人类 Yes	人类 No	Grok Yes	Grok No
否认选举结果	8	2	0	10
气候改变否认	6	4	0	10
未经证实的疫苗主张	9	1	0	10
移民与犯罪叙事	6	4	0	10
大规模枪击事件骗局说法	7	3	0	10

注：每个主题包含 20 条纠正回复，其中由人类发布和由 Grok 介导的纠正各占一半。

图1展示了至少收到一次人身攻击式回应的错误信息纠正所占百分比，并比较了人工直接发布的纠正与 Grok 介导的纠正。在所有五个错误信息主题中，72% 的人工直接发布的纠正都遭到了人身攻击，而 Grok 介导的纠正均未收到人身攻击式回应。

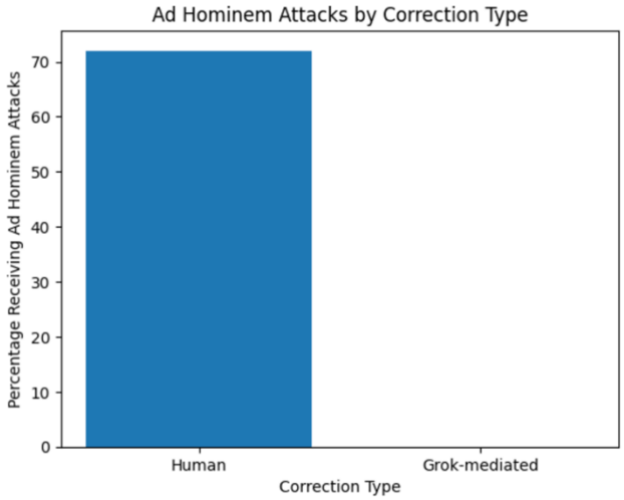


图 1. 按纠正类型展示收到至少一条人身攻击式回应的纠正所占百分比

V. 分析与解读

研究结果直接回应了第二个研究问题，表明在 X 上纠正错误信息时，纠正类型与人身攻击的发生之间存在强关联。在所考察的五个错误信息主题中，人工直接发布的纠正经常遭到人身攻击，而在观察样本中，Grok 介导的纠正没有收到任何人身攻击式回应。这种模式在政治、科学和社会争议领域均一致出现，表明这种关联并非局限于某个单一议题领域或叙事背景。

对这一模式的一种合理解释是，提及 Grok 会改变错误信息纠正过程中人们所感知到的社会动态。当用户通过 Grok 提供纠正性信息时，纠正行为可能会从个人质疑转变为归因于外部系统的媒介化评估。这种转变可能降低人们对人际对抗的感知，进而减少其他用户以个人敌意回应的可能性。相比之下，人工直接纠正会更明确地传递人际分歧的信号。既有研究表明，这可能会加剧冲突，并在在线环境中引发人身攻击式回应。

结果进一步表明，Grok 可能发挥社会中介的作用，而不只是提供信息辅助。通过将纠正性判断归因于人工智能系统，用户似乎重新分配了纠正行为本身的责任，从而可能转移个人责备或敌意。这一解释与有关在线互动的更广泛研究相符；这些研究表明，去人格化、权威归因和媒介化传播会影响受众对争议性主张的回应方式。

关于第一个研究问题，分析证实，在多个错误信息领域都可以观察到Grok介导的纠正，包括否认选举结果、公共卫生错误信息、移民叙事、否认气候变化，以及声称大规模枪击事件是人为策划的说法。然而，由于纠正类型在设计上经过平衡，这些发现并不能估计平台上使用Grok的普遍程度，也不意味着

Grok介导的纠正能够代表X平台上典型用户的行为。

与此同时，不应将这些发现解读为 Grok 天生能够防止敌意，也不应认为用户会刻意调用 Grok 以避免人身攻击。由于本研究依赖观察性数据，因此无法确定用户调用 Grok 的动机，也无法分离出观测到的关联背后的因果机制。语气、措辞、感知到的合法性或受众特征方面的差异，也可能促成了 Grok 介导的纠正中未观察到人身攻击式回应这一现象。

最后，在多种不同错误信息主题中观察到的模式具有一致性，这表明本研究捕捉到的动态可能反映了网络冲突的更广泛特征，而非特定主题的敏感性。如果人工智能介导的纠正在不同领域中系统性地减少人身攻击，这将引发一系列重要问题：与平台整合的人工智能工具如何重塑公共话语、影响用户参与，以及如何影响个人参与错误信息纠正的意愿——尽管这种参与可能面临骚扰风险。

VI. 结论与未来工作

在公共在线空间中纠正错误信息并不仅仅是一项认知任务。这是一种在可见性、评判和风险条件下开展的社会行为。本研究考察了 X 上的用户如何通过调用 Grok 进行错误信息纠正来应对这种风险，以及与人工直接回应相比，这类纠正会受到怎样的回应。在五个高冲突虚假信息领域中，Grok介导的纠正始终与未出现人身攻击相关，而人工直接纠正则经常遭遇充满敌意的个人攻击。

这些发现表明，在错误信息情境中，人工智能系统的价值可能不止于准确性或效率。当用户将纠正性判断归因于平台集成的人工智能时，纠正行为本身似乎会从人际对抗转变为媒介化评估。在这一过程中，Grok 等人工智能系统可能会通过重新分配责任、降低感知到的威胁，以及改变他人对纠正性信息的解读方式，悄然重塑公共分歧中的社会动态。重要的是，本研究并未声称 Grok 能够防止敌意，也未声称用户会有意识地使用人工智能工具来保护自己免受攻击。相反，研究表明，人工智能媒介化纠正与显著不同的社会反应之间存在稳健关联。

这一关联的影响十分重大。公共领域的错误信息纠正，依赖人们愿意冒着遭受反弹的风险发声。如果人工智能媒介化纠正能够降低这种风险，即便只是间接降低，也可能影响参与纠正性话语的人群及其参与频率。与此同时，对人工智能系统的依赖也引发了有关信任、问责以及赋予算法判断的权威性等新问题。要理解这些权衡，就需要超越人工智能系统是否准确这一问题，转而探讨它们在争议性环境中如何发挥社会作用。

未来研究应从多个方向拓展这一研究领域。大规模纵向研究可以评估观察到的模式是否会随时间推移，并在不同平台和文化语境中持续存在。定性分析可以探讨语气、归因和感知到的正当性如何与人工智能媒介化纠正相互作用，进而影响受众回应。实验方法和基于调查的方法可以探究用户诉诸人工智能系统的动机，以及感知到的社交安全感是否会影响纠正行为。最后，对不同人工智能模型和平台设计开展比较研究，可以厘清本文观察到的影响究竟是 Grok 所特有的，还是反映了人工智能介导传播中更广泛的动态。

随着各个平台不断将人工智能系统整合到日常互动中，问题已不再仅仅是这些工具能否帮助识别虚假信息，而是它们能否改变真相受到争议时所处的社会条件。本研究表明，人工智能介导的纠正与显著不同的人际结果相关联，凸显出我们有必要不仅将人工智能视为信息资源，也将其视为在线话语社会生态中的参与者。

数据可用性

本研究期间生成并分析的数据集已在 Kaggle 上公开。该数据集包含纠正级注释，以及分析中所用 X 平台原始公开帖文的直接网址。所有链接均指向数据收集时可公开访问的内容。

参考文献

[1] Caramancion, K. M.（2021年12月）。在线行为反应与假新闻接触及检测准确率之间的关系。载于《2021年IEEE第十二届年度普适计算、电子与移动通信会议（UEMCON）论文集》（第0097-0102页）。IEEE。[2] Caramancion, K. M.（2021年5月）。理解个人观点与言论自由监管之间的关联，以及遭受错误信息/虚假信息误导的风险。载于《2021年IEEE世界人工智能物联网大会（AIIoT）论文集》（第0092-0097页）。IEEE。[3] Del Saz-Rubio, M. M.（2023）。分析西班牙和英国首相在 Twitter 上发布节日问候后引发的回应中与不礼貌相关的语言。《语用学杂志》，206，31-55。[4] Smith-Jones, A. C.（2023）。棍棒和石头或许会打断我的骨头，但言语永远不会伤害我……直到我亲眼看到：关于网络喷子与格赖斯准则及（不）礼貌虚拟言语行为的定性内容分析（博士论文）。利伯蒂大学数字共享平台。https://digitalcommons.liberty.edu/doctoral/4697/[5] Ahn, S. J., Fox, J., Dale, K. R., & Avant, J. A.（2015年）。构建虚拟体验的框架：对环境效能和行为的长期影响。《传播研究》，42（6），839-863。[6] Chen, C., & Shu, K.（2024年）。大语言模型时代打击错误信息：机遇与挑战。《人工智能杂志》，45（3），354-368。[7] Pan, Y., Pan, L., Chen, W., Nakov, P., Kan, M. Y., & Wang, W.（2023年12月）。大语言模型带来的错误信息污染风险。载于《计算语言学协会论文集：EMNLP 2023研究成果》（第1389-1403页）。[8] Zhou, X., Sharma, A., Zhang, A. X., & Althoff, T.（2024年）。使用大语言模型纠正社交媒体上的错误信息。arXiv预印本，arXiv:2403.11169。[9] Lim, G., Tan, B. C. Z., Sim, K. Y. H., Shi, W., Chew, M. H., Hee, M. S., ... & Choo, K. T. W.（2025年）。剑与盾：大语言模型在应对虚假信息中的用途与策略。arXiv预印本，arXiv:2506.07211。[10] Gabriel, S., Lyu, L., Siderius, J., Ghassemi, M., Andreas, J., & Ozdaglar, A.（2024年）。MisinfoEval：另类事实时代的生成式人工智能。arXiv预印本，arXiv:2410.09949。[11] Fu, Y., Foell, S., Xu, X., & Hiniker, A.（2024年5月）。从文本到自我：用户对人工智能媒介化纠正工具在人际传播及自我认知方面的看法。载于《2024年CHI计算机系统人为因素会议论文集》（第1-17页）。[12] Mittelstadt, B., Wachter, S., & Russell, C.（2023年）。为了保护科学，我们必须将大语言模型用作零样本翻译器。《自然·人类行为》，7（11），1830-1832。[13] Williams, A. R.（2024年）。大语言模型愈发具有迷惑性，对选举虚假信息产生重要影响。| 艾伦·图灵研究所。艾伦·图灵研究所。https://www.turing.ac.uk/blog/llms-are-ever-more-convincing-important-consequences-election-disinformation